STATISTICAL METHODS FOR CORRELATED DATA FROM OBSERVATIONAL STUDIES

Hongtao Zhang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2015

Approved by:

Jianwen Cai

Haibo Zhou

Donglin Zeng

David Couper

John Kizer

ABSTRACT

Hongtao Zhang: Statistical Methods for Correlated Data From Observational Studies
(Under the direction of Jianwen Cai and Haibo Zhou)

First, we consider case-cohort studies with multiple disease outcomes. To investigate the effect of a risk factor on different diseases, multiple case-cohort studies are usually conducted. To compare the effect of a risk factor on different types of diseases, times to different disease events need to be modeled simultaneously. Existing case-cohort estimators for multiple disease outcomes utilize only the relevant covariate information in cases and subcohort controls, though many covariates are measured for everyone in the full cohort. Intuitively, making full use of the relevant covariate information can improve efficiency. To this end, we consider a class of doubly-weighted estimators for both regular and generalized case-cohort studies with multiple disease outcomes. The asymptotic properties of the proposed estimators are derived and our simulation studies show that a gain in efficiency can be achieved with a properly chosen weight function. We illustrate the proposed method with a data set from Atherosclerosis Risk in Communities (ARIC) study.

Second, we investigate marginal structural Cox model for clusters of correlated failure time observations. In many studies, subjects in the same community form natural clusters and are thus correlated. We formulate marginal structural Cox model for this type of data and prove the consistency and asymptotic normality of the estimator. Simulation studies show that marginal structural Cox model perform properly by yielding unbiased estimate and satisfactory confidence interval coverage. The proposed method is implemented using a claim data assessing the effectiveness of INSPIRIS home visiting health care program.

Third, we study cluster-based PDS. When the outcome is continuous, the two-stage PDS is an appealing sampling scheme that allows investigators to obtain a more informa-

tive sample. In the Collaborative Perinatal Project (CPP), subjects are clustered within each clinic. Statistical method needs to properly account for cluster-level random effects under PDS scheme. We propose estimation and inference procedures based on a semiparametric profile likelihood function. We show that our estimator is consistent and asymptotically normal. In simulation studies, our cluster-based PDS method provides more efficient estimators compared to linear mixed effect models on an SRS of the same size. We apply the method to CPP data.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In medical studies, multivariate data may occur on various occasions. For example, one subject may experience multiple outcomes of interest. On the other hand, subjects sharing similar characteristics can form intrinsically correlated clusters. Proper methods are needed to analyze such data. This dissertation concentrates on multivariate statistical methods in observation studies, possibly with biased sampling schemes.

Using Full Cohort Information to Improve the Efficiency of Multivariate Marginal Hazard Model for Case-Cohort Studies

As all studies are conducted with a limited budget, the maximum study sizes are often restricted by the cost of the exposure ascertainment. Cost-effective sampling designs have long been desired and play an important role in success of many biological studies.

When the outcome is time-to-event, a popular biased sampling design is case-cohort study. First formally introduced in Prentice (1986), case-cohort design requires a random sample of the full cohort, or 'subcohort'. All subjects of the full cohort are followed until failure or censoring, but complete covariate information is only collected for subjects who experienced failure and for the subjects selected into the subcohort. Case-cohort design is a special form of two-phase sampling design (Breslow and Wellner 2007). In some studies, certain covariates are available on all subjects in the full cohort, while other covariate information that is costly to collect is only assembled among the cases and subcohort controls. The former is referred to as the first-phase covariate data, and the latter as second-phase covariate data. Most case-cohort methods discard the first-phase covariate data in the non-subcohort controls, hence it is intuitive that one may gain efficiency by making full use

of the first-phase covariate data. For example, the Atherosclerosis Risk in Communities (ARIC) study is a large cohort study that involves 15,792 participants. One important aim of ARIC study was to assess lipoprotein-associated phospholipase A_2 (Lp-PLA_2) as potential risk factor of atherosclerosis and its sequelae, so that physicians may consider making Lp-PLA_2 a complementary risk factor beyond the traditional ones. Given the large cohort size and funding limitation, measuring Lp-PLA_2 in labs for all the participants would be infeasible. Alternatively, case-cohort studies were carried out: Lp-PLA_2 were obtained only for patients suffering cardiovascular heart disease (CHD) or stroke, together with a random subcohort that were event-free. Investigators (Ballantyne et al. 2004, 2005) studied candidate biomarkers of inflammation as possible risk factors. However, the first-phase covariate information such as LDL/HDL cholesterol level was not fully utilized. Another feature of ARIC study is that multiple disease (e.g CHD and incident stroke) outcomes were monitored simultaneously. Kang and Cai (2009) proposed a weighted estimating equation approach to fit a marginal proportional hazard model with multiple diseases. Kim et al. (2013) used a modified weight function that was empirically shown to improve the efficiency over Kang and Cai (2009) model. In the first topic, we consider a doubly-weighted approach that utilizes all covariate information to improve the efficiency.

Marginal Structural Cox Models

Randomized clinical trials are generally considered the 'gold standard' in establishing causal relationship due to its ability to balance distributions of subject characteristics across treatment groups. Since the treatment assignment is not confounded with the patient's baseline characteristics, treatment effect can be estimated simply by comparing outcomes between treated and untreated groups.

Due to ethical and other concerns, randomized trials are not always an option. Researchers sometimes rely on observational study designs to investigate the relationship be-

tween exposure and outcome. One major challenge in analyzing data from observational studies is confounding by indication, which is introduced if prognostic factor(s) can be related to both treatment history and outcome. Recent years have seen increasing interests in observational comparative effective research (CER), mainly due to the growing adoption of electronic medical record (EMR) database.

In many observational studies, whether a subject will receive active treatment or not is determined by a number of individual-level prognostic factors such as age and comorbidity. Meanwhile, patients from the same community or clinic form natural clusters, whose members share a similar tendency to be assigned active treatment or otherwise. The INSPIRIS Inc. home visiting provider (HVP) program, for example, was initiated to deliver an intensive program that includes home visits by physicians and nurse practitioners and telephonic case management for a high-risk subset of high risk seniors. It is believed that this HVP program has the potential to increase quality of care and reduce total health care expenditures for elders with chronic conditions. Like other studies, individual's medical history and other factors played an important role in determining the program eligibility. Also, enrollment of HVP program was offered in selected communities in the greater Detroit, Ann Arbor/Lansing and Grand Rapids areas, Michigan. Therefore, subjects living in vicinity form clusters and are potentially correlated. So far, the program has accumulated 1,082 participants and claim data are also available on 10,712 non-participants. First incidence of hospitalization after January 1, 2010, which is the date of initial enrollment of HVP program, is the event of interest. The investigators are interested in whether HVP program can identify health problems at an early stage and increases hospitalizations for preventive treatment. Marginal structural Cox model (Hernan et al. 2000) is useful in analyzing observational data to draw causal inference. The second topic of this dissertation is to investigate the performance of the marginal structural Cox model with clusters and derive its asymptotic properties.

Mixed Effect Model for Probability-dependent Sampling Design

There are numerous situations where the outcome of interest is measured continuously. Outcome-dependent sampling (ODS) scheme (Zhou et al. 2002, Weaver and Zhou 2005) was a two-stage sampling procedure proposed to obtain a more informative sample by over-represent the two distributional tails of X, the exposure that is expensive to ascertain. ODS design is a popular choice for studies that values of outcome Y are known for all subjects, but the exposure variable X may be expensive or difficult to ascertain. The data structure of the ODS sample consists of a first-phase simple random sample (SRS) and a second-phase simple random sample from two tails of Y. ODS scheme is useful when the investigators have some knowledge about the relationship between Y and X. For example, if Y has a linear relationship with X, subjects sampled from two tails of Y distribution are more likely to have X-value that falls in its two distributional tails. However, such prior knowledge is not always feasible. To this end, probability-dependent sampling (PDS) scheme (Zhou et al. 2014) allows investigators to over-sample the two tails of X distribution without having knowledge of X in advance. In some studies, e.g. Collaborative Perinatal Project, participants in the same clinic form natural clusters and are potentially correlated. The third topic of this dissertation is to extend PDS design to cluster-based studies.

In the next chapter, we will review the relevant literature in these areas.

CHAPTER 2: LITERATURE REVIEW

2.1. Marginal Structural Cox Model

2.1.1 Cox Model and Extension to Multivariate Case

Cox proportional hazard model (Cox 1972) is the most popular class of model used to analyze time-to-event data. For n independent subjects, the Cox model can be expressed as:

$$\lambda(t) = \lambda_0(t)exp\{\beta^T X(t)\},$$

where $\lambda_0(t)$ is the unspecified nonparametric baseline hazard function and β is the unknown parameters associated with the vector of possibly time-dependent covariates $X(t)$. Estimation of Cox model is based on partial likelihood (Cox 1975). The maximum partial likelihood estimate (MPLE) $\hat{\beta}$, maximizes the partial likelihood function

$$L(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \{ \frac{e^{\beta^T X_i(t)}}{\sum_{j=1}^{n} Y_j(t) e^{\beta^T X_i(t)}} \}^{\Delta N_i(t)},$$

where $Y(t)$ is the at-risk process and $\Delta N_i(t) = 1$ if the i^{th} subject has event of interest (fails) at time t and 0 otherwise. Andersen and Gill (1982) investigated the asymptotic properties of MPLE $\hat{\beta}$ using martingale theory. Under mild regularity conditions, the MPLE $\hat{\beta}$ is consistent for the true parameter β_0 and is asymptotically normal

$$n^{1/2}(\hat{\beta} - \beta_0) \rightarrow_d N(0, \mathcal{I}_1(\beta_0)^{-1}),$$

in which $\mathcal{I}_1(\beta_0) = \langle n^{-1/2} U(\tau; \beta_0) \rangle$.

In many studies, subjects may be intrinsically correlated. For example, patients that are treated in the same medical clinic form clusters and are likely to be correlated. Another example is that multiple diseases may be monitored simultaneously for one subject. Either example leads to the multivariate survival analysis scenario. Currently, there are in general two classes of models to analyze multivariate time-to-event data. One class is known as the frailty models (Hougaard 1995). We use subscript $k = 1, \ldots, K$ to index the clusters and $i = 1, \ldots, n_0$ to index patients or observations within a cluster. Frailty models have the form

$$\lambda_{ki}(t|Q_k) = Q_k \lambda_0(t) exp\{\beta^T X_{ki}(t)\},$$

in which the cluster-specific random effect Q_k is assumed to follow a known distribution. Q_k resembles a random effect in linear mixed effect models and it is commonly assumed to follow Gamma distribution (Clayton and Cuzick 1985) or positive stable distribution (Hougaard 1986).

The other class is the marginal models. Marginal models leaves the nature of dependence structure completely unspecified, rendering them more robust. Marginal models are usually of greater interest when the dependence structure is not of interest. Wei et al. (1989) suggested fitting a marginal Cox proportional hazard model that can be formulated as

$$\lambda_{ki}(t) = \lambda_{0k}(t) exp\{\beta_k^T X_{ki}(t)\}. \tag{2.1}$$

Each cluster k has a distinct baseline hazard function, which suits well with the multiple disease situation. Model (2.1) can be implemented using standard software packages. Asymptotic results also confirm that the MPLE is a consistent estimator whose variance can be approximated by a robust estimator that is in the sandwich form.

Lee et al. (1992) considered a proportional hazard model with a common baseline

hazard function across all clusters:

$$\lambda_{ki}(t) = \lambda_0(t)exp\{\beta^T X_{ki}(t)\}. \tag{2.2}$$

The asymptotic results were similar to those of model (2.1).

2.1.2 Marginal Structural Cox Model

Marginal structural models (Robins et al. 2000, Hernan et al. 2001) are a class of models used in causal inference. Such models handle the issue of time-dependent confounding in evaluation of the efficacy of interventions by inverse probability weighting for receipt of treatment. Marginal structural Cox model (Hernan et al. 2000) for time-to-event data is an important extension. It takes a form that resembles the regular Cox model, but is formulated using counterfactual arguments. Consider an observational study where the outcome of interest is survival time T. Let $A(t)$ indicate the observed treatment(s) received which can take various forms. For example, it could be an indicator of treatment initiated at baseline, an arbitrary function of dose level, or a time-dependent indicator of treatment received at time t. Let $L(t)$ denote a vector of covariates and $L(0)$ represents baseline covariates. Use overbars to represent history up to time t (t included) such that $\overline{A}(t) = \{A(u) : 0 \leq u \leq t\}$. $\overline{L}(t)$ is defined analogously. Let $\overline{a}$ be any treatment, potentially contrary to what was observed, that a subject could receive. Specifically, $\overline{a} = \{a(t) : 0 \leq t \leq \tau\}$, where τ is the duration of the study. Observed treatment history $\overline{A}(t)$ can be considered a particular realization of $\overline{a}(t)$. There will be a failure time $T_{\overline{a}}$ associated with each possible realization of $\overline{a}$. The simplest case is when we only consider the treatment received at baseline. The counterfactual is thus two-dimensional, with two possible outcomes T_1 and T_0, representing the failure time had the subject been assigned to experimental and control group, respectively.

We need the following three assumptions that are needed for marginal structural models

Assumption 2.1.1. $T = T_{\overline{a}}$ *for any* $\overline{a}$ *such that* $a(t) = A(t), t \leq T$

Assumption 2.1.2. $pr(A(t)|\overline{A}(t^-), \overline{L}(t^-)) > 0$*, for any* $t \in [0, \tau]$ *such that*

$$pr(\overline{A}(t^-), \overline{L}(t^-)) > 0$$

Assumption 2.1.3. $T_{\overline{a}} \perp\!\!\!\perp A(t)|\overline{A}(t^-), \overline{L}(t^-)$*, for any* $\overline{a}$

Assumptions 2.1.1 and 2.1.2 are usually referred to as consistency and positivity assumptions, respectively (Hernan and Robins 2006, Cole and Frangakis 2009). Assumption 2.1.1 states that an individual's observed failure time T is precisely the potential failure time $T_{\overline{a}}$ under a certain observed exposure history $\overline{a}$. Assumption 2.1.2 states that the probability of receiving any particular treatment at time t, given treatment and covariate history up to t, is greater than zero. Assumption 2.1.3 is known as no unmeasured confounding (Hernan et al. 2000). In practice, only assumption 2.1.2 is testable.

We consider the marginal structural Cox model for the hazard of failure at time t had the subject received treatment $\overline{a}$

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) exp\{\beta_0^T a(t)\}, \tag{2.3}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β_0 is the unknown parameter vector. $exp\{\beta_0\}$ will have the interpretation of average treatment hazard ratio.

Estimation of parameters in model (2.3) is carried out via inverse probability weighting technique. Before proceeding to likelihood, we need to discretize the study duration $[0, \tau]$ so that weight functions can be defined. Let $0 \leq t_1 < t_2 < \cdots < t_D \leq \tau$ be D distinct time points, which can be distinct observed times (event or not), or time of scheduled follow-up

visits. Define

$$W(t) = \prod_{t_d \leq t} \frac{1}{pr[A(t_d)|\overline{A}(t_d^-), \overline{L}(t_d)]}. \tag{2.4}$$

At any given time t, the subject is inversely weighted by the probability of receiving the observed history of treatment up to that moment. By inverse probability weighting, we create a hypothetical pseudo-population where $\overline{L}(t) \perp\!\!\!\perp A(t)|\overline{A}(t^-)$ holds at time t.

When $L(t)$ contains confounders that are strongly correlated to treatment $A(t)$, the estimate weight $\hat{W}(t)$ can vary drastically, resulting in high sampling variability in $\hat{\beta}_W$. As a remedy, Robins et al. (2000) and Hernan et al. (2000) suggested using a stabilized version of $W(t)$

$$w(t) = \prod_{t_d \leq t} \frac{pr[A(t_d)|\overline{A}(t_d^-)]}{pr[A(t_d)|\overline{A}(t_d^-), \overline{L}(t_d)]}. \tag{2.5}$$

In (2.5), the excessive contribution of $W(t)$ can be offset by the probability conditional on treatment history solely on the numerator. Using either $W(t)$ or $w(t)$ will result in a weighted partial likelihood function that can be maximized using Newton-Raphson iterative algorithm. In her dissertation, Lee (2013) proved the asymptotic properties of marginal structural Cox model for the case-cohort design when subjects are independent.

2.2. Statistical Methods for Biased Sampling Designs

Because randomized trials are not always an option, researchers sometimes rely on observational study designs to investigate the relationship between outcome and exposure and other covariates. As all studies are conducted with a limited budget, the maximum study sizes are often restricted by the cost of the exposure ascertainment. Cost-effective study designs have long been desired and play an important role in success of many biological studies. Among them, the biased sampling design play an important role. Depending on the nature of outcome of interest, different sampling strategies are utilized. When the

response is intrinsically binary, case-control designs (Prentice and Pyke 1979) are often preferred. The idea of case-control design is to over-sample cases that are believed to be more informative. Ordinary analysis can then be performed on the case-control sample.

On the other hand, there are numerous situations where the outcome of interest is measured continuously. Case-control design cannot be naturally extended to continuous outcome setting. In practice, investigators often dichotomize the outcome based on whether the outcome is above or below a certain cutoff (potentially subjective). However, it is obvious that doing so will discard the information in the continuous outcome. Also, the results may be sensitive to the choice of cutoff. Zhou et al. (2002) proposed an outcome-dependent sampling (ODS) scheme to address this issue. To fix notation, let Y denote the continuous outcome variable and X the vector of covariates. Assume that Y can be partitioned into K mutually exclusive and exhaustive strata by known constants $-\infty = a_0 < a_1 < \cdots < a_{K-1} < a_K = \infty$ and let the kth stratum be represented by $C_k = (a_{k-1}, a_k], k = 1, \ldots, K$. The data structure of the ODS sample consists of an overall simple random sample (SRS) of size n_0 and a simple random sample of size $n_1, \ldots, n_K$ from each of the K strata. The latter is referred to as 'supplemental sample'. Let $n_V = \sum_{k=0}^{K} n_k$ be the total size of the ODS sample and N be the sample size in the population. Borrowing terms from the measurement error literature, we refer to the ODS sample as the 'validation sample', and refer to the rest $n_{\bar{V}} = N - n_V$ observation as the non-validation sample. Let V represent the index set of all observations in the validation sample, and let $\bar{V}$ represent the index set of all observations in the non-validation sample. Their corresponding partitions V_k and $\bar{V}_k$ are also defined. Although response Y is observed for all observations, complete covariate information X is measured only in the validation sample. Zhou et al. (2002) utilized only the validation sample. Let G_X and g_X denote the cumulative distribution and density functions of X, respectively. Also,

$F(u) = pr(Y \leq u)$ and $F(u|x) = pr(Y \leq u|x)$. Then the validation sample likelihood is

$$\begin{aligned} L(\beta, G_X) &= \left\{ \prod_{i=1}^{n_0} f_\beta(y_{0i}|x_{0i}) g_X(x_{0i}) \right\} \times \left[\prod_{k=1}^{K} \prod_{j=1}^{n_k} f_\beta(y_{kj}, x_{kj}|y_{kj} \in C_k) \right] \\ &= \left\{ \prod_{i=1}^{n_0} f_\beta(y_{0i}|x_{0i}) \times \prod_{k=1}^{K} \prod_{j=1}^{n_k} \frac{f_\beta(y_{kj}|x_{kj})}{F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})} \right\} \\ &\times \left\{ \prod_{i=1}^{n_0} g_X(x_{0i}) \times \prod_{k=1}^{K} \prod_{j=1}^{n_k} \frac{g_X(x_{kj})[F(a_k|x_{kj}) - F(a_{k-1}|x_{kj})]}{F(a_k) - F(a_{k-1})} \right\} \\ &= L_1(\beta) \times L_2(\beta, G_X). \end{aligned} \tag{2.6}$$

The second component in (2.6), $L_2(\beta, G_X)$, is a function of β because the conditional distribution function $F(\cdot|x)$ depends upon β. Without specifying G_X, inference about β can be obtained by maximizing $L_1(\beta)$ with respect to β. Zhou et al. (2002) elected to leave G_X unspecified, making it an infinite-dimension nuisance parameter, and used Lagrange multiplier arguments to derive a semiparametric empirical likelihood function. Their discussion focused on partitioning Y into 3 strata and over-sample its two tails C_1, C_3. They profiled $L_2(\beta, G_X)$ by fixing β and obtaining the empirical likelihood function of G_X over all distributions whose support contains the observed X values. During the process, four more parameters were introduced $\eta = (\pi_1, \pi_3, \nu_1, \nu_3)^T$ where $\pi_1 = F(a_1), \pi_3 = \bar{F}(a_3) = 1 - F(a_3)$ and ν_1, ν_3 were Lagrange multipliers. The parameter estimator for $\theta = (\beta^T, \eta^T)^T$ can be obtained iteratively via Newton-Raphson method and its asymptotic properties were derived.

Weaver and Zhou (2005) made attempt to utilize the information in the non-validation sample. It was assumed that, in addition to the complete observations in the ODS, information regarding stratum membership would be retained for the observations in the non-validation sample. Clearly, this assumption would be satisfied as long as continuous response Y is measured for all N observations. Let $f_Y(Y_j; \beta) = \int f_Y(Y_j|x; \beta) dG_X(x)$ be

the unspecified marginal density of Y. The full sample likelihood is proportional to

$$L_F(\beta, G_X) = \left[\prod_{i\in V} f_\beta(Y_i|X_i)\right] \times \left[\prod_{i\in V} dG_X(X_i)\right] \times \left[\prod_{j\in \bar{V}} f_Y(Y_j;\beta)\right]. \tag{2.7}$$

Denote the number of observations in the SRS that belong to the kth stratum as $n_{0,k}$. N_k is defined likewise for the population. Weaver and Zhou (2005) proposed to substitute the unspecified G_X with its consistent empirical estimator, that is,

$$\hat{G}_X(x) = \sum_{k=1}^{K} \frac{N_k}{N} \hat{G}_k(x), \text{where } \hat{G}_k(x) = \sum_{i\in V_k} \frac{I\{X_i \le x\}}{n_k + n_{0,k}}.$$

The resultant unbiased estimator for $f_Y(Y_j;\beta)$ is then

$$\hat{f}_Y(Y_j;\beta) = \sum_{k=1}^{K} \frac{N_k}{N(n_k + n_{0,k})} \sum_{i\in V_k} f_\beta(Y_j|X_i). \tag{2.8}$$

Substituting (2.8) into (2.7) and applying the log-transformation, one can obtain the estimated log-likelihood

$$\hat{l}_F(\beta) = \left[\sum_{i\in V} log f_\beta(Y_i|X_i)\right] + \left[\sum_{j\in \bar{V}} log\left\{\sum_{k=1}^{K} \frac{N_k}{N(n_k + n_{0,k})} \sum_{i\in V_k} f_\beta(Y_j|X_i)\right\}\right].$$

Consistent estimator is again found using Newton-Raphson method.

In practice, there may exist an auxiliary variable W, which is available for all N observations, for the exposure X. It is thus necessary to incorporate the information implied by W into the statistical analysis. Let Z be the covariates which are observed for all subjects. It is assumed that Z is related to the conditional density of Y. On the other hand,

W provides no additional information about Y when X and Z are known. Zhou et al. (2011b) considered an outcome-auxiliary-dependent sampling (OADS) scheme. In addition to the K partitions of Y, assume that W can be partitioned into J mutually exclusive and exhaustive strata by known constants $-\infty = b_0 < b_1 < \cdots < b_{J-1} < b_J = \infty$ and let the jth stratum be represented by $B_j = (b_{j-1}, b_j], j = 1, \ldots, J$. Then the population can be partitioned into $T = K \times J$ strata on the domain of $Y \times W$. For notational simplicity, use $\Delta_t, t = 1, \ldots, T$ to denote the strata. $V, \bar{V}$ and corresponding partitions $V_t, \bar{V}_t$ are defined analogously. The outcome-auxiliary-dependent sample also consists of two components: an overall SRS and a supplemental sample of size n_t from tth stratum. These two components make up the validation sample, whose complement is referred to as non-validation sample. Using similar arguments as in Weaver and Zhou (2005), the full sample with likelihood

$$L_F(\beta) = \left[\prod_{t=0}^{T}\prod_{i\in V_t} f_\beta(Y_i|Z_i, X_i)dG(X_i|Z_i, W_i)\right] \times \left[\prod_{t=0}^{T}\prod_{i\in \bar{V}_t}\int_X f_\beta(Y_i|Z_i, x)dG(x|Z_i, W_i)\right]. \tag{2.9}$$

$G(X|Z, W)$ is again estimated non-parametrically. Let S denote the $d-$dimensional information components of (Z, W) in the sense that $G(X|Z, W) = G(X|S)$ almost surely. Zhou et al. (2011b) proposed to estimate $G(x|s)$ via kernel smoothing. Specifically,

$$\hat{G}(x|s) = \sum_{t=1}^{T} \hat{\pi}_t(s)\hat{G}_t(x|s),$$

where

$$\hat{\pi}_t(s) = \frac{\sum_{i=1}^{N} I\{(Y_i, W_i) \in \Delta_t\}\phi_h(S_i - s)}{\sum_{i=1}^{N} \phi_h(S_i - s)}, \hat{G}_t(x|s) = \frac{\sum_{i\in V_t} I\{X_i \le x\phi_h(S_i - s)\}}{\sum_{i\in V_t} \phi_h(S_i - s)}.$$

$\phi_h(\cdot)$ here is a $d-$dimensional kernel function with bandwidth h. Substituting $\hat{G}(x|s)$ into (2.9) to obtain the estimated likelihood function, then one can employ Newton-Raphson method to obtain the estimate of β and consequently make inferences. In some studies, observations may be intrinsically clustered (e.g. within medical clinics) and statistical methods need to account for the cluster-level random effect. Xu and Zhou (2012) investigated cluster-based OADS. They postulated a linear mixed effect model for $f_\beta(y|x,z)$

$$Y_{mti} = \beta_0 + \beta_1 X_{mti} + \beta_2 Z_{mti} + u_m + e_{mti} \tag{2.10}$$

The additional subscript $m = 1, \ldots, M$ indexes the clusters. $u_m \sim N(0, \sigma_u^2)$, $e_{mti} \sim N(0, \sigma^2)$ and $u_m \perp\!\!\!\perp e_{mti}$. Their likelihood function had a similar form as in Zhou et al. (2011b). However, numerical integration technique was implemented to address the complexity introduced by the cluster-level random effect u_m. Wang and Zhou (2010) considered the OADS design with categorical response variable. Estimation and inference were carried out based on an estimated likelihood function.

So far in this section, the mean model linking the response to the exposure of interest X and covariates Z was assumed to be linear. Specifically, the linear model is $E(Y|X,Z) = \beta^T X + \gamma^T Z$, where β and γ are unknown regression parameters. In many studies, it is desirable to make such relationship flexible. Zhou et al. (2011a) and Qin and Zhou (2011) proposed a partial linear model leaving the functional form of exposure X unspecified. On the other hand, Zhou et al. (2011c) employed a similar partial linear model, but left the functional form of covariate Z unspecified. Other variations of the original ODS design were explored. Ding et al. (2012) considered a special situation in which the exposure variables are fully unobservable but only the summation of them can be observed. This type of data is usually encountered in genetic studies. ODS design can also be implemented in studies where the response is time-to-event (Ding et al. 2014).

Recently, a probability-dependent sampling (PDS) scheme has been proposed (Zhou et al. 2014). Assume that the domain of the exposure X is partitioned into three mutually exclusive intervals: $(-\infty, x_L] \cup (x_L, x_U] \cup (x_U, \infty)$. Like ODS, an SRS is drawn from the population at the first stage. Before supplemental sampling, a model for $E(X|Y, Z)$ is fitted using the first-phase SRS. On the basis of this model, the chances of a new subject's X conditional on $Y = y$ and $Z = z$, will be in $(-\infty, x_L]$ and (x_U, ∞) are predicted by $\hat{\phi}_1(y, z) = \hat{pr}(X < x_L|Y, Z)$ and $\hat{\phi}_3(y, z) = \hat{pr}(X > x_U|Y, Z)$ respectively. Then supplemental samples are drawn from those whose X are more likely to fall on the distributional tails of X. For example, random samples can be drawn from those with $\hat{\phi}_1(y, z) > c_1$ or with $\hat{\phi}_3(y, z) > c_3$, where c_1 and c_3 are thresholds. Assume that $f_\beta(y|x, z)$ has the linear form

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + e,$$

where e is normal random error. Let $G(X, Z)$ and $g(X, Z)$ denote the joint CDF and PDF of (X, Z). Define

$$\begin{aligned}\pi_k &= pr\{\phi_k(Y, Z) \geq c_k\} \\ &= \int\int\int f_\beta(Y|X, Z)g(X, Z)I\{(Y, Z) : \phi_k(Y, Z) \geq c_k\}dY dX dZ.\end{aligned}$$

The log-likelihood for the validation sample is

$$\begin{aligned}l(\beta, \{p_i\}, \pi_1, \pi_3) &= \left\{\sum_{i\in V} log f_\beta(Y_i|X_i, Z_i)\right\} + \left\{\sum_{i\in V} log(p_i) - n_1 log(\pi_1) - n_3 log(\pi_3)\right\} \\ &= l_1(\beta) + l_2(\beta, \{p_i\}, \pi_1, \pi_3), \qquad (2.11)\end{aligned}$$

where $p_i = g(X_i, Z_i)$. Using Lagrange multiplier arguments, one can profiled l_2 in (2.11) over $\{p_i\}$ by fixing (β, π_1, π_3) and obtaining the empirical likelihood function of $\{p_i\}$ over all distributions whose support contains the observed values of X and Z. The estimate

β that maximizes the resulting semiparametric empirical log-likelihood is found using Newton-Raphson algorithm. In their simulation study, PDS estimator was shown to have improved efficiency over the ODS estimator.

We have so far focused on reviewing statistical methods for biased sampling when response Y is completely continuously measured. When the outcome of interest is time-to-event, subject to censoring, an important biased sampling design is case-cohort sampling. We review the related literature in the next section.

2.3. Statistical Methods for Case-cohort Design

2.3.1 Univariate Case-cohort Design

As an alternative design to reduce cost and achieve the same goal as a full cohort study, case-cohort design was first formally introduced in Prentice (1986). The design involves the collection of covariate data for all cases in the full study cohort and for a small random sample of size $\tilde{n}$ from the entire cohort called the subcohort, denoted by C. The subcohort in a given stratum constitutes the comparison set of cases occurring at a range of failure times. The subcohort also provides a basis for covariate monitoring during the course of cohort follow-up. The relative risk regression model considered had the form

$$\lambda\{t : Z(u), 0 \leq u < t\} = \lambda_0(t) r\{\beta^T Z(t)\}, \tag{2.12}$$

where $r(x)$ is a fixed function satisfying $r(0) = 1$. The resultant pseudo likelihood took the form

$$\tilde{L}(\beta) = \prod_{i=1}^{n} (r_{ii} / \sum_{l \in \tilde{R}(t_i)} r_{li})^{\Delta_i}, \tag{2.13}$$

where $r_{li} = Y_l(t_i)r\{\beta^T Z_l(t_i)\}$. Assuming no tied times, the risk set $\tilde{R}(t_i)$ contains all the at-risk subjects in the subcohort at time t_i, and the subject who experienced uncensored failure at t_i. Since (2.13) does not generally possess a partial likelihood interpretation, it was termed pseudo likelihood. The maximum pseudo likelihood estimate $\tilde{\beta}_P$ satisfies $U(\tilde{\beta}_P) = 0$, where the pseudo likelihood estimating function

$$U(\beta) = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} \Delta_i (c_{ii} - \sum_{l \in \tilde{R}(t_i)} b_{li} / \sum_{l \in \tilde{R}(t_i)} r_{li}) \tag{2.14}$$

in which $b_{li} = Y_l(t_i)Z_l(t_i)r'\{\beta^T Z_l(t_i)\}, c_{li} = b_{li}(r\{\beta' Z_l(t_i)\})^{-1}$ and $r'(u) = dr(u)/du$. Under some mild regularity conditions, $n^{-1/2}U(\beta_0)$ was shown to converge to a normal variate with mean 0 and variance matrix $\boldsymbol{A}$, where

$$\boldsymbol{A} = \sum_{i=1}^{n} [var\{U_i(\beta)\} + 2 \sum_{\{k|t_k<t_i\}} cov\{U_k(\beta), U_i(\beta)\}].$$

Consequently, $n^{-1/2}(\tilde{\beta}_P - \beta_0)$ was shown to converge in distribution to a zero mean normal distribution with a sandwich type variance matrix $\boldsymbol{S} = \boldsymbol{\Omega}^{-1}\boldsymbol{A}\boldsymbol{\Omega}^{-1}$, which can be consistently estimated by $n\boldsymbol{I}(\tilde{\beta}_P)^{-1}\tilde{\boldsymbol{V}}(\tilde{\beta}_P)\boldsymbol{I}(\tilde{\beta}_P)^{-1}$ where

$$\tilde{\boldsymbol{V}}(\beta) = \sum_{j=1}^{n} \Delta_j \{v_{jj} + 2\tilde{\Delta}(t_j) \sum_{\{k|t_k<t_j\}} \Delta_k v_{kj}\},$$

with

$$\begin{aligned} v_{kj} &= -\sum_{i \in \tilde{R}(t_j)} (\frac{B_k + b_{jk} - b_{ik}}{R_k + r_{jk} - r_{ik}})'(c_{ij} - \frac{B_j}{R_j}) r_{ij} R_j^{-1}, \\ R_j &= \sum_{l \in \tilde{R}(t_j)} r_{lj}, B_j = \sum_{l \in \tilde{R}(t_j)} b_{lj}, \end{aligned}$$

and $\tilde{\Delta}(t) = 1$ if $\tilde{R}(t) \neq C$ and 0 otherwise.

A natural estimator of the cumulative baseline hazard function can be written

$$\hat{\Lambda}_0(t) = \tilde{n}n^{-1} \int_0^t \left[\sum_{l \in C} Y_l(w) r\{\tilde{\beta_P}^T Z_l(w)\} \right]^{-1} d\bar{N}(w), \tag{2.15}$$

where $\bar{N} = N_1 + \cdots + N_n$. It was shown that the results could be extended to a stratified design in which a baseline covariate is used to partition the entire cohort into Q strata and stratum-specific relative risk regression models are specified.

Self and Prentice (1988) developed the asymptotic theory for the case-cohort maximum pseudo likelihood estimator and related quantities using a combination of martingale and finite population convergence results. They considered an estimator, denoted by $\tilde{\beta}_{SP}$, that maximizes the pseudo likelihood function very similar to (2.13). The pseudo log-likelihood can be written

$$log\tilde{L}(\beta, t) = \sum_{i \in C} \int_0^t log(r\{\beta^T Z_i(u)\}) dN_i(u) - \int_0^t log\left[\sum_{l \in C} Y_l(u) r\{\beta^T Z_i(u)\} \right] d\bar{N}(u). \tag{2.16}$$

Under mild regularity conditions plus additional assumptions regarding the stability of subcohort averages, $n^{-1/2}\tilde{U}(\beta)$ was shown to converge to a zero mean normal random variable whose covariance matrix being $\boldsymbol{\Sigma}(\beta_0) = \boldsymbol{D}(\beta_0) + \boldsymbol{A}(\beta_0)$ in which $\boldsymbol{D}(\beta_0)$ can be consistently estimated by the information matrix generated by the pseudo log-likelihood (2.16) and the matrix $\boldsymbol{A}(\beta_0)$ reflects the efficiency loss induced by the sampling of the subcohort. Following Taylor expansion arguments, $n^{1/2}(\tilde{\beta}_{SP} - \beta_0)$ was shown to converge in distribution to a zero mean Gaussian random variable with covariance matrix $\boldsymbol{D}(\beta_0)^{-1} + \boldsymbol{D}(\beta_0)^{-1}\boldsymbol{A}(\beta_0)\boldsymbol{D}(\beta_0)^{-1}$. An estimate of the cumulative baseline hazard function $\Lambda_0(t)$ has the same form of (2.15), except for substituting $\tilde{\beta}_P$ with $\tilde{\beta}_{SP}$.

The only difference between $\tilde{\beta}_P$ and $\tilde{\beta}_{SP}$ is the construction of the comparison risk set

$\tilde{R}(t)$: the risk set in Prentice (1986) includes all subcohort members at risk at t plus any individuals who are not in the subcohort but who experienced failure at t, while the risk set in Self and Prentice (1988) only contains members in the subcohort. As a result, the two asymptotic covariance matrices are slightly different. However, it was shown that $\tilde{\beta}_P$ and $\tilde{\beta}_{SP}$ are asymptotically equivalent provided an individual's contributions to $S^{(1)}$ and $S^{(0)}$ are asymptotically negligible. The covariance matrix associated with $\tilde{\beta}_P$ was shown to converge to $\boldsymbol{D}(\beta_0)^{-1} + \boldsymbol{D}(\beta_0)^{-1}\boldsymbol{A}(\beta_0)\boldsymbol{D}(\beta_0)^{-1}$.

Both variance estimators proposed by Prentice (1986) and Self and Prentice (1988) have complicated form. The application of case-cohort design was hindered partly due to the perceived difficulty in variance computation. Attempts were made to address this issue. Wacholder et al. (1989) proposed (1) a variance estimator based on superpopulation under the null hypothesis $H_0 : \beta = 0$ (2) a variance estimator obtained from a modified bootstrap resampling procedure. However, the drawbacks of the two estimators are obvious: the former is only valid under the null hypothesis, while the latter can be highly computational intensive when dealing with large studies. It was noted that the former estimator may be useful in predicting the power of the case-cohort design, and its efficiency, compared to the full cohort study.

A robust variance estimator was proposed in Barlow (1994), based on the influence of an individual observation on the overall score. Barlow's estimator bypassed the explicit estimation of $\boldsymbol{A}(\beta_0)$ in Self and Prentice (1988) and was shown to be a jackknife variance estimator. Specifically, he considered a weighted version of pseudo likelihood function: the conditional probability of failure at time t_j is given by

$$p_i(t_j) = \frac{Y_i(t_j)w_i(t_j)r_i(t_j)}{\sum_{k=1}^{n} Y_k(t_j)w_k(t_j)r_k(t_j)},$$

where $r_i(t) = exp\{\beta^T Z_i(t)\}$ and the weight $w_i(t)$ reflecting subcohort membership and

current failure status is defined as

$$w_i(t) = \begin{cases} 1 & \text{if } dN_i(t) = 1, \\ m(t)/\tilde{m}(t) & \text{if } dN_i(t) = 0 \text{ and } i \in C, \\ 0 & \text{if } dN_i(t) = 0 \text{ and } i \notin C. \end{cases}$$

$m(t)$ is the number of disease-free individuals in the cohort at risk at time t and $\tilde{m}(t)$ is the number disease-free in the subcohort at time t. The time-invariant estimator of $m(t)/\tilde{m}(t)$ is given by $n/\tilde{n}$. Estimation of β_0 follows directly from the pseudo log-likelihood function $\sum_t \sum_i log(p_i(t))dN_i(t)$. The estimator $\tilde{\beta}_B$ maximizing the pseudo log-likelihood has a robust variance estimator given by $\hat{var}(\tilde{\beta}_B) = n^{-1} \sum_{i=1}^{n} \hat{e}_i \hat{e}_i^T$ where $\hat{e}_i = \tilde{\beta}_B - \tilde{\beta}_{B(-i)}$ is the change in $\tilde{\beta}_B$ if the ith individual were deleted. To estimate $\hat{e}_i$, let $c_i(t)$ denote the influence of an individual observation on the overall score for person i at time t, we have

$$c_i(t) = \int_0^t Y_i(u)[dN_i(u) - \lambda_i(u)][Z_i(u) - E(u)]d\bar{N}(u),$$

where $E(t)$ is the conditional expectation of the covariate at time t. Then $\hat{e}_i$ can be approximated by $I^{-1}(\tilde{\beta}_B)\hat{c}_i(t)$, with $I^{-1}(\beta)$ being the inverse of the information matrix generated by the pseudo log-likelihood and

$$\hat{c}_i(t) = \int_0^t Y_i(u)[dN_i(u) - \hat{p}_i(u)][Z_i(u) - \hat{E}(u)]d\bar{N}(u).$$

$\hat{p}_i(u), \hat{E}(u)$ are corresponding estimates of $p_i(u), E(u)$ by substituting β with $\tilde{\beta}_B$.

Lin and Ying (1993) provided a general solution to the problem of missing covariate data under the Cox regression model. Case-cohort design can be treated as a special case under their framework. Suppose that the data consist of i.i.d. random quintuplets

$\{X_i, \Delta_i, Z_i(\cdot), H_{0i}(\cdot), \boldsymbol{H}_i(\cdot)\}$, where $Z_i(\cdot)$ is a p-dimensional covariate vector that may not be completely observed. $H_{0i}(t)$ is the subcohort indicator which equals 1 if the ith subject is in the subcohort at time t. $\boldsymbol{H}_i(t)$ is a $p \times p$ diagonal matrix with indicator functions $\{H_{1i}(t), \ldots, H_{pi}(t)\}$ as diagonal elements, where $H_{ji} = 1$ if $Z_{ji}(t)$ is observed and $H_{ji} = 0$ otherwise $(j = 1, \ldots, p)$. $H_{0i}(t)$ determines whether or not the ith subject is included in the estimation of $\bar{Z}(\beta, t)$ and $H_{ji}(t)$ indicates whether or not the ith subject contributes directly to the jth component of the estimating function. Under MCAR assumption, the approximate partial likelihood score function for estimating β_0 can be written as

$$\tilde{U}_H(\beta) = \sum_{i=1}^{n} \Delta_i \boldsymbol{H}_i(X_i)\{Z_i(X_i) - Z_H(\beta, X_i)\}$$

where $Z_H(\beta, t) = S_H^{(1)}(\beta, t)/S_H^{(0)}(\beta, t)$ and

$$S_H^{(d)} = n^{-1} \sum_{i=1}^{n} H_{0i}(t) Y_i(t) exp\{\beta^T Z_i(t)\} Z_i(t)^{\otimes d}.$$

The APLE $\tilde{\beta}_H$ is the root to $\tilde{U}_H(\beta) = 0$. In addition to the common regularity conditions, two more are required to derive the asymptotic properties of $\tilde{\beta}_H$:

(A) All covariates have bounded total variations, that is, $\int_0^\infty |dZ_{ji}(t)| + |Z_{ji}(0)| \leq K$ for some $K > 0$ and all i, j

(B) There exist $k_0 > 0$ and $\eta_0 > 0$ such that for $j = 0, 1, \ldots, p$ and $r = 0, 1$,

$$\sup_{|d| \leq n^{-k_0}} \left[n^{-1} \left| \sum_{i=1}^{n} \{H_{ji}(t) - H_{ji}(t+d)\} \right| + |h_j(t) - h_j(t+d)| \right] = o_p(n^{-(1/2)-\eta_0})$$

and

$$\sup_{|d|\leq n^{-k_0}} \left[n^{-1} \left\| \sum_{i=1}^{n} \{ \boldsymbol{Z}_i(t) - \boldsymbol{Z}_i(t+d) \} \right\| + \| s^{(r)}(\beta_0, t) - s^{(r)}(\beta_0, t+d) \| \right]$$
$$= o_p(n^{-(1/2)-\eta_0}),$$

where $s^{(r)}$ are the limits of $S_H^{(r)}(\beta, t)$.

Then $n^{1/2}(\tilde{\beta}_H - \beta_0)$ was shown to follow a normal distribution with mean 0 and sandwich type covariance matrix $\boldsymbol{A}(\beta_0)^{-1}\boldsymbol{B}(\beta_0)\boldsymbol{A}(\beta_0)^{-1}$ where

$$\begin{aligned}
\boldsymbol{A}(\beta) = & \lim_{n\to\infty} -n^{-1}\frac{\partial \tilde{U}_H(\beta)}{\partial \beta}, \\
\boldsymbol{B}(\beta) = & E\{\boldsymbol{W}_1(\beta)^{\otimes 2}\}, \\
\boldsymbol{W}_i(\beta) = & \Delta_i \boldsymbol{H}_i(X_i)\{Z_i(X_i) - z_H(\beta, X_i)\} \\
- & \int_0^{X_i} \{\boldsymbol{h}(t)/h_0(t)\} H_{0i}(t) exp\{\beta^T Z_i(t)\}\{Z_i(t) - z_H(\beta, t)\}\lambda_0(t)dt,
\end{aligned}$$

and $z_H(\beta, t) = s_H^{(1)}(\beta, t)/s_H^{(0)}(\beta, t)$, $s_H^{(d)} = E[S_H^{(0)}]$, $\boldsymbol{h}(t) = E[\boldsymbol{H}_i(t)]$ and $h_j(t) = E[H_{j1}(t)]$.

The proposed framework incorporates many sampling designs. In particular, for case-cohort designs with time-independent covariates, it is clear that $(H_{i1}(t), \ldots, H_{ip}(t))$ are time-invariant and $\boldsymbol{h}(t) \equiv (h)$. The covariance estimator is much easier to compute than those of Prentice (1986) and Self and Prentice (1988), especially in the presence of time-dependent covariates. Furthermore, incomplete covariate measurements on the cases are allowed. The estimator of the cumulative baseline hazard function is given by $\hat{\Lambda}(\tilde{\beta}_H, t)$

$$\hat{\Lambda}(\beta, t) = \sum_{i=1}^{n} \frac{I(X_i \leq t)\Delta_i H_{0i}(X_i)}{n S_H^{(0)}(\tilde{\beta}_H, X_i)}.$$

Chen and Lo (1999) derived a class of estimating equations for case-cohort sampling, each depending on a different estimator of the population distribution, which lead naturally to simple estimators that improve on pseudo likelihood estimator of Prentice (1986). Their key idea that enables an improvement is that, in constructing the risk set to be used in estimating equations, the information in all case samples should be completely rather than only partially utilized. The pseudo likelihood estimating equation is given by

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[z_i(t) - \frac{\sum_{j \in C \cup \{i\}} Y_j(t) z_j(t) e^{\beta^T z_j(t)}}{\sum_{j \in C \cup \{i\}} Y_j(t) e^{\beta^T z_j(t)}} \right] dN_i(t) = 0 \tag{2.17}$$

The second term in (2.17) estimates $m(t)$, the conditional mean of Z given $X = t$ and $\Delta = 1$, through the identity

$$\begin{aligned} E[Z|X = t, \Delta = 1] &= \frac{E(Ze^{\beta^T Z} I_{(X \geq t)})}{E(e^{\beta^T Z} I_{(X \geq t)})} \\ &= \frac{pE(Ze^{\beta^T Z} I_{(X \geq t)}|\Delta = 1) + (1-p)E(Ze^{\beta^T Z} I_{(X \geq t)}|\Delta = 0)}{pE(e^{\beta^T Z} I_{(X \geq t)}|\Delta = 1) + (1-p)E(e^{\beta^T Z} I_{(X \geq t)}|\Delta = 0)} \end{aligned} \tag{2.18}$$

where $p = pr(\Delta = 1)$. Based on (2.18), they derived a class of estimating equations that yield different case-cohort estimators. Let $N(n)$ be the size of the cohort (subcohort), $N_1(n_1)$ and $N_0(n_0)$ be the numbers of cases and controls in the cohort (subcohort), respectively. Also, use $C^1(\tilde{C}^1)$ and $C^0(\tilde{C}^0)$ to denote, respectively, the index sets of all cases and all controls in the cohort (subcohort). The subscript t restricts an index set to individuals at risk time t, so that the at risk indicator $Y_j(t)$ is no longer needed in (2.18). Depending on whether the full cohort is well-defined and whether failure probability p is known, they used different strategies to estimate (2.18) and substitute it in (2.17). Specificly, they considered the following scenarios:

Case 1. let $\hat{n_1}/n$ be the estimator of p. Then $\hat{\beta}_1$ solves $U_1(\beta) = 0$ where $U_1(\beta)$ is

$$\sum_{i=1}^{n} \int_0^\tau \left[z_i(t) - \frac{\{n_1/(nN_1)\} \sum_{j \in C_t^1} z_j(t) e^{\beta^T z_j(t)} + (1/n) \sum_{j \in \tilde{C}_t^0} z_j(t) e^{\beta^T z_j(t)}}{\{n_1/(nN_1)\} \sum_{j \in C_t^1} e^{\beta^T z_j(t)} + (1/n) \sum_{j \in \tilde{C}_t^0} e^{\beta^T z_j(t)}} \right] dN_i(t)$$

Case 2. If the full cohort is well defined so that N, N_1, N_0 are known. Substitue p with a better estimator $\hat{p} = N_1/N$. The resultant $\hat{\beta}_2$ solves $U_2(\beta) = 0$ where $U_2(\beta)$ is

$$\sum_{i=1}^{n} \int_0^\tau \left[z_i(t) - \frac{(1/N) \sum_{j \in C_t^1} z_j(t) e^{\beta^T z_j(t)} + \{N_0/(Nn_0)\} \sum_{j \in \tilde{C}_t^0} z_j(t) e^{\beta^T z_j(t)}}{(1/N) \sum_{j \in C_t^1} e^{\beta^T z_j(t)} + \{N_0/(Nn_0)\} \sum_{j \in \tilde{C}_t^0} e^{\beta^T z_j(t)}} \right] dN_i(t)$$

Case 3. If the population case percentage p is known. Then $\hat{\beta}_3$ solves $U_3(\beta) = 0$ where $U_3(\beta)$ is

$$\sum_{i=1}^{n} \int_0^\tau \left[z_i(t) - \frac{(p/N_1) \sum_{j \in C_t^1} z_j(t) e^{\beta^T z_j(t)} + \{(1-p)/n_0\} \sum_{j \in \tilde{C}_t^0} z_j(t) e^{\beta^T z_j(t)}}{(p/N_1) \sum_{j \in C_t^1} e^{\beta^T z_j(t)} + \{(1-p)/n_0\} \sum_{j \in \tilde{C}_t^0} e^{\beta^T z_j(t)}} \right] dN_i(t)$$

Chen and Lo (1999) proved that $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are all consistent and their respective asymptotic variances were derived. Their limited simulation study showed that $\hat{\beta}_2$ performed the best in the sense that it yielded the smallest standard error.

Borgan et al. (2000) considered stratified case-cohort sampling designs and proposed three estimator based on different ways to construct risk sets and to estimate inverse sampling proportion $w_k(t)$, where $k = 1, \ldots, K$ indexes strata in the full cohort based on a stratum variable that is available for everyone. Their work was extended by Samuelsen et al. (2007), who considered stratified generalized case-cohort design with surrogate variable that are predictive of the main exposure variables.

In the aim of improving the efficiency, Kulich and Lin (2004) proposed a class of

weighted estimating equations under stratified case-cohort design. They considered the situation when some covariates (usually covariates that are less expensive to collect, e.g., age and blood type) are available for the full cohort, while other costly covariates (e.g., genotype) are only in the case-cohort sample. The former was termed first-phase covariate data and the latter second-phase covariate data. The idea was to improve the efficiency by making fuller use of the first-phase covariate data.

Consider a cohort of n subjects who can be divided into K mutually exclusive strata based on a discrete stratum variable V, which is available for everyone in the full cohort. Let $\alpha_k = pr(\xi = 1|V = k), (k = 1, \ldots, K)$ where ξ is the subcohort indicator. Let n_k be the number of subjects in the kth stratum and let $q_k \equiv pr(V = k)$. They proposed so-called doubly weighted estimator so that a separate set of (time dependent) weights is used for each covariate in Cox model to estimate the sampling proportion. In specific, they defined $\boldsymbol{A}_{ik}(t)$, subject to certain regularity conditions, as a diagonal matrix with m potentially different random processes on the diagonal, where m is the number of covariates in the target Cox model with common baseline hazard. Then the quantity analogous to the subcohort sampling probabilities is given by

$$\hat{\boldsymbol{\alpha}}_k(t) = \left\{ \sum_{i=1}^{n_k} (1 - \Delta_{ki}) \boldsymbol{A}_{ik}(t) \right\}^{-1} \left\{ \sum_{i=1}^{n_k} \xi_{ki} (1 - \Delta_{ki}) \boldsymbol{A}_{ki}(t) \right\}.$$

Accordingly, the time-dependent weight function has the form

$$\boldsymbol{\varrho}_{ki}(t) = \Delta_{ki} \boldsymbol{I}_m + (1 - \Delta_{ki}) \xi_{ki} \hat{\boldsymbol{\alpha}}_k(t)^{-1}.$$

Then the at-risk covariate average is estimated by:

$$\bar{\boldsymbol{Z}}_{KL}(\beta, t) \equiv \left\{ S_{KL}^{(0)}(\beta, t) \right\}^{-1} S_{KL}^{(1)}(\beta, t)$$

where $S^{(d)}_{KL}(\beta,t) \equiv n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\boldsymbol{\varrho}_{ki}(t)Y_{ki}(t)exp\{\beta^T Z_{ki}(t)\}Z_{ki}(t)^{\otimes d}$. Note $S^{(1)}_{KL}(\beta,t)$ is an m-vector, whereas $S^{(0)}_{KL}(\beta,t)$ is a diagonal $m\times m$ matrix. The pseudo score function is defined as

$$U_{KL}(\beta)=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^{\tau}\{Z_{ki}(t)-\bar{\boldsymbol{Z}}_{KL}(\beta,t)\}dN_{ki}(t). \tag{2.19}$$

Their estimator, $\hat{\beta}_{KL}$, solves the equation $U_{KL}(\beta)=0$. It was shown that

$$\sqrt{n}(\hat{\beta}_{KL}-\beta_0)\longrightarrow_d N(0,\mathcal{I}_F^{-1}+\mathcal{I}_F^{-1}\boldsymbol{\Sigma}_{KL}\mathcal{I}_F^{-1}).$$

Like many of its counterparts in other case-cohort estimators, the variance estimator consists of two parts: the variance of full-data partial likelihood estimator plus the efficiency loss due to case-cohort sampling. $\mathcal{I}_F$ is the limiting information matrix generated by (2.19).

It is not intuitive to estimate the cumulative baseline hazard function $\Lambda_0(t)$ since the quantity $S^{(0)}_{KL}(\beta,t)$ is not a scalar when dealing with more than one covariates in the Cox model. Alternatively, Kulich and Lin (2004) proposed to estimate $\Lambda_0(t)$ using an existing method, that is,

$$\hat{\Lambda}_0(t)=n^{-1}\int_0^t\{S^{(0)}_{\Lambda}(u,\hat{\beta}_{\Lambda})\}^{-1}\sum_{k,i}dN_{ki}(u),$$

in which the footnote Λ indicates that the quantity comes from an existing method (Kulich and Lin (2004) used estimator II in Borgan et al. (2000), denoted as $\hat{\beta}_B$).

Through arguments in semiparametric efficiency, they claimed the optimal form of second level weight $\boldsymbol{A}_{ki}(t)$ is given by

$$\boldsymbol{A}_{ki}(t)\equiv diag[\{\hat{Z}_{ki}-\bar{Z}_B(t,\hat{\beta}_B)\}exp\{\hat{\beta}_B^T\hat{Z}_{ki}\}Y_{ki}(t)] \tag{2.20}$$

where $\hat{Z}_{ki}$ equals Z_{ki} if the subject is selected in the case-cohort sample. Otherwise, the missing part in Z_{ki} is imputed from a rich model regressing the missing covariate(s) on

all observed covariates. It was pointed out that $\hat{\beta}_{KL}$ may not always perform well in finite sample sizes. Hence, the authors also developed a combined doubly weighted estimator $\hat{\beta}_{CDW}$ and derived its asymptotic properties. In their simulation, they considered the case when a surrogate of the missing covariate is available.

2.3.2 Multivariate Case-cohort Design

Despite the progress in the methods for univariate case-cohort studies, there is a very limited collection of literature addressing the analysis of case-cohort data with multiple disease outcomes. Lu and Shih (2006) focused their discussion on large study cohort with many clusters for which the investigators seek to evaluate the effect of risk factors or the effect of an intervention program at a population average level. Assume that the full cohort consists of n independent clusters, and each cluster contains m_i correlated individuals, for $i = 1, \ldots, n$. It is assumed that the individuals within the same cluster are exchangeable conditional on covariates. Due to the clustered feature of the full cohort, modified case-cohort sampling procedures are needed. Lu and Shih (2006) proposed three designs corresponding to various scenarios in the full cohort:

Design A: suppose that a complete roster of clusters is available for the full cohort, then randomly sample r individuals per cluster, and collect covariate data from the selected individuals and all failures from the entire cohort.

Design B: when the full cohort is large and enumeration of every cluster is impossible, then randomly sample $\tilde{n}$ clusters without replacement and collect covariate data from all members of each selected cluster and all failures from the entire cohort.

Design C: In *design B*, if the cluster size is large, combine the principle of *design A*.

Their approach was built upon the work by Lee et al. (1992), in which a common

cumulative baseline hazard function $\Lambda_0(t)$ and common regression coefficients β_0 were assumed. Estimation and asymptotic inference under design A is different from those under design B and C, due to the fact that sampling $\tilde{n}$ clusters without replacement induces correlation among the $\tilde{n}$ clusters. Let H_i be the cluster indicator that takes value 1 if cluster i belongs to the subcohort, and 0 otherwise, and η_{ij} be the individual indicator that equals 1 if individual (i, j) is selected in the subcohort, and 0 otherwise. Under design A, the estimator $\hat{\beta}$ solves the estimating equation $U_A(\beta) = 0$ where

$$U_A(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Big[Z_{ij}(X_{ij}) - \bar{E}(\beta, X_{ij}) \Big] \Delta_{ij}.$$

Here,

$$\begin{aligned} \bar{E}(\beta, u) &= \bar{S}^{(1)}(\beta, u)/\bar{S}^{(0)}(\beta, u), \\ \bar{S}^{(d)}(\beta, u) &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \eta_{ij} Y_{ij}(u) exp\{\beta^T Z_{ij}(u)\} Z_{ij}(u)^{\otimes d}. \end{aligned}$$

Under mild regulatory conditions, $\hat{\beta}$ was shown to be consistent and asymptotically normal. The variance of $n^{1/2}(\hat{\beta} - \beta_0)$ is given by $\boldsymbol{A}(\beta_0)^{-1} \boldsymbol{\Omega}(\beta_0) \boldsymbol{A}(\beta_0)^{-1}$, where $\boldsymbol{\Omega}(\beta_0) = E\{\boldsymbol{W}_1(\beta_0)^{\otimes 2}\}$,

$$\begin{aligned} \boldsymbol{W}_1(\beta) = \sum_{j=1}^{m} \int_0^{\tau} & \{Z_{ij}(u) - \bar{z}(\beta, u)\} \\ & \times [dN_{ij}(u) - \eta_{ij} Y_{ij}(u) exp\{\beta^T Z_{ij}(u)\}/\{\alpha_r s^{(0)}(\beta, u)\} dF(u)], \\ \alpha_r = \frac{r}{m}. & \end{aligned}$$

Under designs B and C, the corresponding estimating equation is very similar to that under

design A, except that $\bar{S}^{(d)}(\beta, u)$ is defined as

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} H_i\eta_{ij}Y_{ij}(u)exp\{\beta^T Z_{ij}(u)\}Z_{ij}(u)^{\otimes d}$$

With additional conditions, the estimator $\tilde{\beta}$ was also consistent, asymptotically normal and the variance is $\boldsymbol{A}(\beta_0)^{-1}\tilde{\boldsymbol{\Omega}}(\beta_0)\boldsymbol{A}(\beta_0)^{-1}$. Since $\tilde{\Omega}(\beta)$ has a complicated expression, the authors also proposed alternative ways to estimate it, including a modified bootstrap resampling procedure based on Wacholder et al. (1989). Through simulation studies, they concluded that the statistical efficiency is improved when sampling greater $\tilde{n}$ clusters and/or more individuals per cluster (r).

Unlike case-cohort design with clusters, for multiple disease outcomes, the common baseline assumption is not realistic and could lead to biased estimates if the baseline hazard function is indeed different for different disease outcomes. The work by Kang and Cai (2009) addressed this issue by considering a marginal disease-specific Cox proportional hazard model

$$\lambda_{ik}(t) = Y_{ik}(t)\lambda_{0k}(t)e^{\beta_0^T Z_{ik}(t)} \tag{2.21}$$

in which $k = 1, \ldots, K$ denotes different diseases and $i = 1, \ldots, n$ denotes subjects in the full cohort. They also extended their estimation and inference procedure for generalized case-cohort designs with multiple outcomes. The design is in effect a two-stage sampling procedure. First, select a subcohort of fixed size $\tilde{n}$ from the cohort by simple random sampling. Use ξ_i to denote the subcohort indicator that equals 1 if subject i is included in the subcohort. After the sampling of a subcohort, subsequent samplings of cases outside the subcohort follow. For the kth disease, sample a fixed number of $\tilde{n}_{c,k}$ cases that are outside the subcohort by simple random sampling without replacement. Let η_{ik} be the indicator for the ith subject outside the subcohort with the kth disease being selected into the sample. Define $\tilde{\alpha} = pr(\xi_i = 1) = \tilde{n}/n$ and $\tilde{q}_k = pr(\eta_{ik} = 1|\Delta_{ik} = 1, \xi_i = 0) =$

$\tilde{n}_{c,k}/(n_k - \tilde{n}_k)$, where $n_k(\tilde{n}_k)$ denotes the number of the kth disease cases in the cohort (subcohort). The potentially time-varying weight function is given by

$$w_{ik}(t) = \Delta_{ik}\xi_i + (1 - \Delta_{ik})\xi_i\hat{\alpha}_k(t)^{-1} + \Delta_{ik}(1 - \xi_i)\eta_{ik}\hat{q}_k(t)^{-1} \tag{2.22}$$

in which

$$\hat{\alpha}_k(t) = \frac{\sum_{i=1}^{n}(1 - \Delta_{ik})\xi_i Y_{ik}(t)}{\sum_{i=1}^{n}(1 - \Delta_{ik})Y_{ik}(t)},$$
$$\hat{q}_k(t) = \frac{\sum_{i=1}^{n}\Delta_{ik}(1 - \xi_i)\eta_{ik}Y_{ik}(t)}{\sum_{i=1}^{n}\Delta_{ik}(1 - \xi_i)Y_{ik}(t)}.$$

Therefore, the weighted estimating equation has the form

$$U_{KC}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} w_{ik}(t)\{Z_{ik}(t) - \bar{Z}_k(\beta, t)\}dN_{ik}(t). \tag{2.23}$$

Note that, in original case-cohort design, $w_{ik}(t)$ is not required in (2.23) since all cases are sampled and each has $w_{ik}(t) \equiv 1$. For generalized case-cohort design, however, it cannot be omitted due to the case sampling. Denote the resultant estimator as $\hat{\beta}_{KC}$, which was shown to be consistent and asymptotically normal. The variance-covariance matrix of $n^{1/2}(\hat{\beta}_{KC} - \beta_0)$ is given by $\boldsymbol{A}(\beta_0)^{-1}\boldsymbol{\Sigma}(\beta_0)\boldsymbol{A}(\beta_0)^{-1}$ in which

$$\boldsymbol{\Sigma}(\beta_0) = Q(\beta_0) + \frac{1 - \tilde{\alpha}}{\tilde{\alpha}}V^I(\beta_0) + (1 - \tilde{\alpha})\sum_k pr(\Delta_{1k} = 1)\frac{1 - \tilde{q}_k}{\tilde{q}_k}V_k^{II}(\beta_0), \tag{2.24}$$

where

$$
\begin{aligned}
Q(\beta_0) &= E[\sum_k M_{\bar{z},1k}(\beta_0)]^{\otimes 2}, \\
V^I(\beta_0) &= E[\sum_k (1-\Delta_{1k}) \\
&\quad \cdot \int_0^\tau [R_{1k}(\beta_0,t) - \mu_k(t)^{-1} Y_{1k}(t) E[(1-\Delta_{1k}) R_{1k}(\beta_0,t)]\}] d\Lambda_{0k}(t)]^{\otimes 2}, \\
V_k^{II}(\beta_0) &= E\{[M_{\bar{z},1k}(\beta_0) - \int_0^\tau \theta_k(t)^{-1} Y_{1k}(t) E[\Delta_{1k} dM_{\bar{z},1k}(\beta_0)]]^{\otimes 2} | \Delta_{1k}=1, \xi_1=0\}.
\end{aligned}
$$

$\boldsymbol{A}(\beta_0)$ is the information matrix generated by (2.23). Also,

$$
\begin{aligned}
\tilde{Z}_{ik}(\beta,t) &= Z_{ik}(t) - \bar{z}_k(\beta,t), \\
M_{ik}(\beta,t) &= N_{ik}(t) - \int_0^t Y_{ik}(u) e^{\beta^T Z_{ik}(u)} d\Lambda_{0k}(u), \\
M_{\tilde{z},ik}(\beta) &= \int_0^\tau \tilde{Z}_{ik}(\beta,t) dM_{ik}(t), \\
R_{ik}(\beta,t) &= Y_{ik}(t) \tilde{Z}_{ik}(\beta,t) e^{\beta^T Z_{ik}(u)}, \\
\mu_k(t) &= E\{(1-\Delta_{1k}) Y_{1k}(t)\}, \\
\theta_k(t) &= E\{Y_{1k}(t) | \Delta_{1k}=1\}.
\end{aligned}
$$

The baseline cumulative hazard function Λ_{0k} can be consistently estimated by

$$
\hat{\Lambda}_{0k}(\hat{\beta}_{KC},t) = \int_0^\tau \frac{\sum_{i=1}^n dN_{ik}(t)}{n\hat{S}^{(0)}(\hat{\beta}_{KC},u)}.
$$

The variance-covariance matrix is estimated by plugging in consistent empirical estimators of corresponding components. Kim et al. (2013) attempted to improve the efficiency over Kang and Cai (2009) by making use of information on all outcomes. Their weighted

estimator used the weight function

$$\psi_{ik}(t) = \left\{1 - \prod_{j=1}^{K}(1-\Delta_{ij})\right\} + \prod_{j=1}^{K}(1-\Delta_{ij})\xi_i\tilde{\alpha}_k^{-1}(t),$$

where $\tilde{\alpha}_k(t) = \sum_{i=1}^{n}\xi_i\{\prod_{j=1}^{K}(1-\Delta_{ij})\}Y_{ik}(t)/\sum_{i=1}^{n}\{\prod_{j=1}^{K}(1-\Delta_{ij})\}Y_{ik}(t)$. Their estimator was shown to be consistent and asymptotically normal, with a sandwich-type variance. Simulation studies confirmed the efficiency gain in finite sample.

CHAPTER 3: MORE EFFICIENT CASE-COHORT ESTIMATORS

3.1. Introduction

Case-cohort design is widely used in large cohort studies when it is prohibitively costly to assemble covariate history for all subjects in the full cohort. First introduced in Prentice (1986), case-cohort design requires a random sample in the full cohort, or 'subcohort'. All subjects in the full cohort are followed until failure or censoring occurs, but complete covariate information is only collected for subjects who experienced failure and for those subjects selected into the subcohort. Case-cohort design is a special form of two-phase sampling design (Breslow and Wellner 2007).

For data from case-cohort studies for a single disease outcome, many methods have been proposed under the Cox proportional hazard model (Cox 1972) framework. Prentice (1986) and Self and Prentice (1988) studied a pseudo-likelihood approach, which modified the partial likelihood (Cox 1975) by weighting the contributions of cases and subcohort controls differently. Barlow (1994) provided an easier alternative approach to compute the asymptotic variance. Chen and Lo (1999) used a refined procedure to estimate the at-risk average to achieve efficiency gain. Borgan et al. (2000) considered a stratified case-cohort design and used time-varying weights based on the at-risk process to improve the efficiency of the parameter estimates. When it is of interest to compare the effect of a risk factor on different diseases, marginal models are appealing. Despite the advances in methods for univariate case-cohort designs, literature on the marginal models for case-cohort data with multiple disease outcomes is scarce. Kang and Cai (2009) proposed a weighted estimating equation approach to fit a marginal proportional hazard model with multiple diseases. Kim et al. (2013) proposed a modified weight function that used all available disease status to

improve efficiency. Both of these methods only used the covariate information collected on cases and subjects in the subcohort.

In many studies, certain covariates are available on all subjects in the full cohort, while other covariate information that is costly to collect is only assembled among the cases and subjects in the subcohort. The former is referred to as the first-phase covariate data, and the latter as second-phase covariate data. For example, the Atherosclerosis Risk in Communities (ARIC) study is a large cohort study that involved 15,792 participants. One important aim of ARIC study was to assess lipoprotein-associated phospholipase A_2 (Lp-PLA_2) as potential risk factor of atherosclerosis and its sequelae, so that physicians may consider making Lp-PLA_2 a complementary risk factor beyond the traditional ones. Given the large cohort size and funding limitation, measuring Lp-PLA_2 in labs for all the participants would be infeasible. Alternatively, case-cohort studies were carried out: Lp-PLA_2 were obtained only for participants suffering cardiovascular heart disease (CHD) or stroke, together with a subcohort that were free of CHD or stroke (Ballantyne et al. 2004, 2005). Lp-PLA_2 is thus the second-phase covariate and the first-phase covariates are the information collected on the full cohort at the cohort visit, such as race, gender, lipid measurements, etc. To compare the effect of Lp-PLA_2 on the incident stroke and CHD, the two disease outcomes need to be modeled simultaneously to properly account for their correlation. The methods proposed by Kang and Cai (2009) and Kim et al. (2013) can be applied in this situation. However, only covariate information collected on the cases and subjects in the subcohort are used. It is desirable to use relevant covariate information collected on the full cohort to improve efficiency. For a single survival outcome, Kulich and Lin (2004) proposed a doubly-weighted estimator that used all available first-phase covariate data and postulated a regression model for second-phase covariate(s) on first-phase covariate(s). However, with multiple diseases, to our knowledge, no work has been done to fully utilize the first-phase covariates. In this paper, we aim to investigate a doubly-weighted approach to improve efficiency with multiple diseases with data from multiple traditional

case-cohort studies. Furthermore, we will also consider generalized case-cohort designs. Generalized case-cohort designs are usually conducted when the disease is not rare, but there is limited resources in biospecimen. Under such situation, instead of taking all the cases, a random sample of cases outside the subcohort will be drawn (Cai and Zeng 2007, Kang and Cai 2009). It will be of interest to examine the doubly-weighted approach for the generalized case-cohort studies.

In this paper, we focus on the analysis of time-to-event data with multiple disease outcomes and consider a doubly-weighted approach with the aim of improving the efficiency under (generalized) case-cohort design. Section 3.2 formulates the doubly-weighted estimating equation framework. Asymptotic properties are presented in section 3.3. In section 3.4 we report the simulation results. ARIC study was analyzed in section 3.5. We give some concluding remarks in section 3.6.

3.2. Model and Estimation

3.2.1 Notations and Model Definition

Suppose that there are n independent subjects in the full cohort and K disease outcomes of interest. Consider independent vectors of potential failure times $T_i = (T_{i1}, \ldots, T_{iK})^T$, $i = 1, \ldots, n, k = 1, \ldots, K$. Similarly, we use $C_i = (C_{i1}, \ldots, C_{iK})^T$ to denote the potential right censoring time vectors. In practice, it is common to have $C_{i1} = \ldots C_{iK} = C_i$. The observed time $X_{ik} = T_{ik} \wedge C_{ik}$. Let $\Delta_{ik} = I(T_{ik} \leq C_{ik})$ denote the event indicator, $N_{ik}(t) = I(X_{ik} \leq t, \Delta_{ik} = 1)$ the counting process, and $Y_{ik}(t) = I(X_{ik} \geq t)$ the at-risk process for disease k of subject i, respectively. Let $Z_{ik}(t)$ be a $p \times 1$ potentially time-dependent covariate vector that can be decomposed into two components: a $p_1 \times 1$ vector of first-phase covariates $V_{ik}(t)$, and a $p_2 \times 1$ vector of second-phase covariates $W_{ik}(t)$. The time-dependent covariates are assumed to be 'external' in the sense that they are not

affected by the outcome processes (Kalbfleisch and Prentice 2002). We assemble all the covariates into a vector $Z_i = (Z_{i1}, \ldots, Z_{iK})^T$. Finally, τ is the study end time.

Suppose that potential failure time T_{ik} arises from a Cox-type proportional marginal hazards model (Cai and Prentice 1995)

$$\lambda_{ik}(t|Z_{ik}(t)) = Y_{ik}(t)\lambda_{0k}(t)e^{\beta_0^T Z_{ik}(t)}, \tag{3.1}$$

where $\lambda_{0k}(t)$ is the unspecified, disease-specific baseline hazard function and β_0 is a $p \times 1$ vector of fixed and unknown regression parameters. Disease-specific covariate effects can be accommodated by defining $\beta^* = (\beta_1^T, \ldots, \beta_K^T)^T$ and $Z_{ik}(t)^* = (0_{i1}^T, \ldots, Z_{ik}(t)^T, \ldots, 0_{iK}^T)$ where β_k denotes the disease-k-specific effect for covariate $Z_{ik}(t)$, $k = 1, ..., K$. Under the two systems of notation, we have $\beta_k^T Z_{ik}(t) = \beta^{*T} Z_{ik}(t)^*$.

3.2.2 Estimation

If the data were complete, for $d = 0, 1, 2$, define $S_{k,F}^{(d)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_{ik}(t) Z_{ik}(t)^{\otimes d} e^{\beta^T Z_{ik}(t)}$, with $a^{\otimes 0} = 1, a^{\otimes 1} = a, a^{\otimes 2} = aa^T$. The relative risk parameter β_0 can be estimated by solving the pseudo partial likelihood score equation

$$U_F(\beta) = \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \{Z_{ik}(t) - \bar{Z}_{k,F}(\beta, t)\} dN_{ik}(t) = 0, \tag{3.2}$$

where $\bar{Z}_{k,F}(\beta, t) = S_{k,F}^{(1)}(\beta, t)/S_{k,F}^{(0)}(\beta, t)$. Under the case-cohort design, (3.2) cannot be calculated because covariate vector $Z_{ik}(t)$ is not fully observed for subjects that are neither in the subcohort nor among sampled cases. Instead, we consider a weighted version of pseudo likelihood score function in which information from a completely observed subject represents multi-fold information from potentially missing subjects.

Assume that we sample without replacement to obtain a subcohort of size $\tilde{n}$. Subcohort sampling is followed by the sampling of non-subcohort cases, that is, for disease k, we sample m_k subjects without replacement from cases that are outside the subcohort. Let ξ_i be an indicator of subcohort membership which equals 1 if subject i is sampled into the subcohort and 0 otherwise. Similarly, we define η_{ik} as the indicator for the ith subject outside the subcohort with the kth disease being selected into the sample. For any i, the subcohort sampling probability $\tilde{\alpha} = Pr(\xi_i = 1) = \tilde{n}/n$ and disease-specific case sampling probability $\tilde{q}_k = Pr(\eta_{ik} = 1|\Delta_{ik} = 1, \xi_i = 0) = m_k/(n_k - \tilde{n}_k)$, where n_k and $\tilde{n}_k$ denote the number of cases for the kth disease in the cohort and in the subcohort, respectively.

Marginal proportional hazards model for case-cohort studies with multiple disease outcomes was first investigated by Kang and Cai (2009), who embedded the at-risk processes in estimating $\tilde{\alpha}$ and $\tilde{q}_k$. The motivation of using the doubly-weighted estimator arises from the intuition that one could incorporate additional information beyond the at-risk processes, hence, obtain a more efficient estimator. Further, it is desirable to have the flexibility of weighting each covariate in (3.1) differently, which could lead to improved precision. We hereafter use the superscript/subscript 'KC' and 'DW' to indicate that the quantity, function or estimate is obtained from implementing the $\tilde{\beta}_{II}$ estimator in Kang and Cai (2009) and our doubly-weighted estimator, respectively.

Let

$$\tilde{w}_{ik}(t) = \Delta_{ik}\xi_i I_p + (1 - \Delta_{ik})\xi_i \hat{\alpha}_k(t)^{-1} + \Delta_{ik}(1 - \xi_i)\eta_{ik}\hat{q}_k(t)^{-1},$$

where

$$\hat{\alpha}_k(t) = \{\sum_{i=1}^{n}(1 - \Delta_{ik})A_{ik}(t)\}^{-1}\{\sum_{i=1}^{n}(1 - \Delta_{ik})\xi_i A_{ik}(t)\}, \tag{3.3}$$

and

$$\hat{q}_k(t) = \{\sum_{i=1}^{n}\Delta_{ik}(1 - \xi_i)B_{ik}(t)\}^{-1}\{\sum_{i=1}^{n}\Delta_{ik}(1 - \xi_i)\eta_{ik}B_{ik}(t)\}, \tag{3.4}$$

where $A_{ik}(t)$ and $B_{ik}(t)$ denote diagonal matrices with p potentially different random processes on their respective diagonals. Each of the p covariates in model (3.1) can have its dedicated process to estimate the subcohort sampling probability $\hat{\alpha}_{k,l}(t)$ and case sampling probability $\hat{q}_{k,l}(t), l = 1, \ldots, p$. Define $S^{(d)}_{k,DW}(\beta, t) = n^{-1} \sum_{i=1}^{n} \tilde{w}_{ik}(t) Y_{ik}(t) Z_{ik}(t)^{\otimes d} e^{\beta^T Z_{ik}(t)}, d = 0, 1, 2$, and the at-risk average process $\bar{Z}_{k,DW}(\beta, t) = \{S^{(0)}_{k,DW}(\beta, t)\}^{-1}\{S^{(1)}_{k,DW}(\beta, t)\}$. We propose to obtain $\hat{\beta}_{DW}$ by solving a doubly-weighted score function:

$$U_{DW}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^{\tau} \tilde{w}_{ik}(t)\{Z_{ik}(t) - \bar{Z}_{k,DW}(\beta, t)\} dN_{ik}(t) = 0. \tag{3.5}$$

Unlike other weighting schemes where weights and $S^{(0)}_k(\beta, t)$ are scalar functions, both $S^{(0)}_{k,DW}(\beta, t)$ and $\tilde{w}_{ik}(t)$ in the doubly-weighted estimating equation in (3.5) are $p \times p$ diagonal matrices. The second level weights, $A_{ik}(t)$ and $B_{ik}(t)$ in (3.3) and (3.4), are diagonal matrices with p potentially different random processes on their respective diagonals. Kang and Cai (2009) estimator is a special case of the doubly-weighted estimator class by setting both $A_{ik}(t)$ and $B_{ik}(t)$ to $Y_{ik}(t) \cdot I_p$. The estimator in Kim et al. (2013) also belongs to this class by setting $A_{ik}(t) = \{\prod_{j=1}^{K}(1 - \Delta_{ij})\}Y_{ik}(t)$ and $B_{ik}(t)$ is not applicable under traditional case-cohort study. Another choice of second level weight is similar to the 'optimal' weight proposed in Kulich and Lin (2004). It was a $p \times p$ diagonal matrix in the form of

$$A_{ik}(t) = diag\left[\{\hat{Z}_{ik}(t) - \bar{Z}_{k,KC}(\hat{\beta}_{KC}, t)\} exp\{\hat{\beta}^T_{KC}\hat{Z}_{ik}(t)\} Y_{ik}(t)\right], \tag{3.6}$$

where $\hat{\beta}_{KC}$ and $\bar{Z}_{k,KC}(\hat{\beta}_{KC}, t)$ were parameter estimate and estimated at-risk average process obtained from implementing the Kang and Cai (2009) model. $\hat{Z}_{ik} = (\hat{Z}_{ik,1}, \ldots, \hat{Z}_{ik,p})^T$ was a p-vector of observed or estimated covariates. Calculating weight (3.6) requires another consistent and asymptotically normal multivariate case-cohort estimator. Other consistent estimators for the two quantities can also be used.

Doubly-weighted estimator $\hat{\beta}_{DW}$ can be obtained via Newton-Raphson algorithm by iteratively solving (3.5) until convergence criterion is met. Specifically, the estimator in the step $k+1$ is $\beta_{DW}^{(k+1)} = \beta_{DW}^{(k)} - D_{DW}(\beta_{DW}^{(k)})^{-1} U_{DW}(\beta_{DW}^{(k)})$, where $D_{DW}(\beta)$ is the derivative of $U_{DW}(\beta)$ with respect to β. Due to the matrix nature of $S_{k,DW}^{(0)}(\beta, t)$, special attention is needed to compute $D_{DW}(\beta)$. Explicit form of $D_{DW}(\beta)$ is given in section 3.7.

We propose to use a Breslow-Aalen type estimator for the baseline cumulative hazard function $\Lambda_{0k}(t)$. The form of the estimator is the same as the one proposed in Kang and Cai (2009) with the estimator for β replaced by $\hat{\beta}_{DW}$. Specifically,

$$\hat{\Lambda}_{0k}(\hat{\beta}_{DW}, t) = \int_0^t \frac{\sum_{j=1}^n \rho_{jk}(u) dN_{jk}(u)}{n S_{k,KC}^{(0)}(\hat{\beta}_{DW}, u)},$$

where

$$\rho_{jk}(u) = \Delta_{ik}\xi_i + (1-\Delta_{ik})\xi_i \hat{\alpha}_k^{KC}(u)^{-1} + \Delta_{ik}(1-\xi_i)\eta_{ik}\hat{q}_k^{KC}(u)^{-1},$$

$$\alpha_k^{KC}(u) = \{\sum_{i=1}^n (1-\Delta_{ik}) Y_{ik}(u)\}^{-1} \{\sum_{i=1}^n (1-\Delta_{ik})\xi_i Y_{ik}(u)\},$$

$$q_k^{KC}(u) = \{\sum_{i=1}^n \Delta_{ik}(1-\xi_i) Y_{ik}(u)\}^{-1} \{\sum_{i=1}^n \Delta_{ik}(1-\xi_i)\eta_{ik} Y_{ik}(u)\},$$

and $S_{k,KC}^{(0)}(\beta, u) = n^{-1} \sum_{i=1}^n \rho_{ik}(u) Y_{ik}(u) e^{\beta^T Z_{ik}(u)}$ are the scalar functions used in Kang and Cai (2009). Based on the results in Kang and Cai (2009), this estimator is consistent and converges weakly to a zero mean Gaussian process if $\hat{\beta}_{DW}$ is a consistent estimator of β_0. We will establish the consistency of $\hat{\beta}_{DW}$ in the next section.

3.3. Asymptotic Properties of General Doubly Weighted Estimator

3.3.1 Asymptotic Results

We present the asymptotic properties of the doubly-weighted estimator. For $k = 1, \ldots, K$, define the following limiting quantities:

$$s_k^{(d)}(\beta, t) = E\{S_{k,F}^{(d)}(\beta, t)\}(d = 0, 1, 2), \bar{z}_k(\beta, t) = s_k^{(1)}(\beta, t)/s_k^{(0)}(\beta, t),$$

$$v_k(\beta, t) = \frac{s_k^{(2)}(\beta, t)s_k^{(0)}(\beta, t) - s_k^{(1)}(\beta, t)^{\otimes 2}}{s_k^{(0)}(\beta, t)^2}, G_k(\beta) = \int_0^\tau v_k(\beta, t)s_k^{(0)}(\beta, t)d\Lambda_{0k}(t).$$

We assume the usual regularity conditions, as required in Spiekerman and Lin (1998):

Assumption 3.3.1. *$(T_i, C_i, Z_i), i = 1, \ldots, n$ are independent and identically distributed*

Assumption 3.3.2. *$pr\{Y_{ik}(t) = 1\} > 0$ for $t \in [0, \tau], i = 1, \ldots, n$ and $k = 1, \ldots, K$*

Assumption 3.3.3. *$|Z_{ik}(0)| + \int_0^\tau |dZ_{ik}(t)| < D_z < \infty$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ almost surely, where D_z is a constant*

Assumption 3.3.4. *$G_k(\beta_0)$ is positive definite for $k = 1, \ldots, K$*

Assumption 3.3.5. *(Finite interval) $\int_0^\tau \lambda_{0k}(t)dt < \infty$ for $k = 1, \ldots, K$*

Assumption 3.3.6. *(Asymptotic stability) There exists a neighborhood $\mathcal{B}$ of β_0 such that*

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}} \|S_{k,F}^{(d)}(\beta, t) - s_k^{(d)}(\beta, t)\| \to_p 0$$

for $d = 0, 1, 2$ and $k = 1, \ldots, K$

Assumption 3.3.7. *(Asymptotic regularity) For all $\beta \in \mathcal{B}$ and $k = 1, \ldots, K$: $s_k^{(1)}(\beta, t) = \frac{\partial}{\partial\beta}s_k^{(0)}(\beta, t)$, $s_k^{(2)}(\beta, t) = \frac{\partial^2}{\partial\beta\partial\beta^T}s_k^{(0)}(\beta, t)$ where $s_k^{(0)}(\cdot, t), s_k^{(1)}(\cdot, t), s_k^{(2)}(\cdot, t)$ are continuous functions of $\beta \in \mathcal{B}$, uniformly in $t \in [0, \tau]$, $s_k^{(0)}(\cdot, t)$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$*

Assumption 3.3.8. *(Lindeberg condition) There exists a $\delta > 0$ such that as $n \to \infty$*

$$n^{-1/2} \sup_{i,k,t} \|Z_{ik}(t)\| Y_{ik}(t) I\{\beta_0^T Z_{ik}(t) > -\delta \|Z_{ik}(t)\|\} \to_p 0$$

We also need the following conditions concerning case-cohort samples and second level weights:

Assumption 3.3.9. *(Nontrivial subcohort and case sampling) As $n \to \infty$, $\tilde{\alpha}$ converges to a constant on $(0, 1]$; similarly, for $k = 1, \ldots, K$, $\tilde{q}_k$ converges to a constant on $(0, 1]$*

Assumption 3.3.10. *For each component $Z_{ik,l}(t)$ of $Z_{ik}(t)$, $var \int_0^\tau |dV_{ik,l}(t)| < \infty$, where $V_{ik,l}(t) = Z_{ik,l}(t) exp\{\beta_0^T Z_{ik}(t)\}$. For each diagonal element $A_{ik,l}(t)$ of $A_{ik}(t)$, $var \int_0^\tau |dA_{ik,l}(t)| < \infty$. Diagonal elements of $B_{ik}(t)$ require a similar condition.*

Assumption 3.3.11. *$A_{ik}(t)$ is independent of ξ_i, and $B_{ik}(t)$ is independent of η_{ik}, for $k = 1, \ldots, K$*

Assumption 3.3.12. *the absolute values of the diagonal elements of $\mu_k(t) \equiv E_k[(1 - \Delta_{1k}) A_{1k}(t)]$ and $\theta_k(t) \equiv E_k[\Delta_{1k} B_{1k}(t)]$ are bounded away from 0 for all $t \in [0, \tau]$*

Assumption 3.3.12 is required in order to prove the asymptotic properties of $\hat{\alpha}_k(t)$ and $\hat{q}_k(t)$. As long as the elements on the diagonal of $A_{ik}(t)$ or $B_{ik}(t)$ are nonnegative (e.g., $Y_{ik}(t)$), this condition is trivial. However, this assumption may not hold if we use the weight function (3.6). We relax this condition in next section. This will enable us to use arbitrary second level weights.

We present the asymptotic results here and provide the outline of the proof in section 3.8. Define

$$M_{ik}(t) = N_{ik}(t) - \int_0^t Y_{ik}(u) e^{\beta_0^T Z_{ik}(u)} d\Lambda_{0k}(u), \quad \tilde{Z}_{ik}(\beta, t) = Z_{ik}(t) - \bar{z}_k(\beta, t),$$

$$M_{\bar{z},ik}(\beta) = \int_0^\tau \tilde{Z}_{ik}(\beta,t)dM_{ik}(t), \quad R_{ik}(\beta,t) = Y_{ik}(t)\tilde{Z}_{ik}(\beta,t)e^{\beta^T Z_{ik}(u)}.$$

Asymptotic properties of $\hat{\beta}_{DW}$ are summarized in the following theorem:

Theorem 3.3.1. *(Asymptotic properties of $\hat{\beta}_{DW}$)*

Under conditions 3.3.1-3.3.12, $\hat{\beta}_{DW}$ solving the estimating equation $U_{DW}(\hat{\beta}_{DW}) = 0$ is a consistent estimator of β_0 and

$$\sqrt{n}(\hat{\beta}_{DW} - \beta_0) \to_d N(0, G(\beta_0)^{-1}\Sigma(\beta_0)G(\beta_0)^{-1}),$$

where $G(\beta) = \sum_k G_k(\beta)$ and

$$\Sigma(\beta_0) = Q(\beta_0) + \frac{1-\tilde{\alpha}}{\tilde{\alpha}}V^I(\beta_0) + (1-\tilde{\alpha})\sum_k pr(\Delta_{1k}=1)\frac{1-\tilde{q}_k}{\tilde{q}_k}V_k^{II}(\beta_0), \qquad (3.7)$$

where

$$\begin{aligned}
Q(\beta_0) &= E\Big\{\sum_k M_{\bar{z},1k}(\beta_0)\Big\}^{\otimes 2}, \\
V^I(\beta_0) &= var\Big\{\sum_k (1-\Delta_{1k})\int_0^\tau \{R_{1k}(\beta_0,t) - \mu_k(t)^{-1}A_{1k}(t)E[(1-\Delta_{1k})R_{1k}(\beta_0,t)]\}d\Lambda_{0k}(t)\Big\}, \\
V_k^{II}(\beta_0) &= var\Big\{M_{\bar{z},1k}(\beta_0) - \int_0^\tau \theta_k(t)^{-1}B_{1k}(t)E[\Delta_{1k}dM_{\bar{z},1k}(\beta_0,t)]\Big|\Delta_{1k}=1,\xi_1=0\Big\}.
\end{aligned}$$

The asymptotic variance of $\hat{\beta}_{DW}$ has three components: the variance of the full data, the variation due to subcohort sampling, and the variation due to further case sampling if generalized case-cohort design is conducted. Unknown quantities can be estimated by substituting proper consistent estimators for their theoretical counterparts. See section 3.8 for details.

3.3.2 Generalization to Arbitrary Second Level Weight

In this section, we relax assumption 3.3.12, which will enable us to use arbitrary second level weights. For notational simplicity, we drop the subscript l by assuming $p = 1$. For $p \geq 2$, the operation is on each diagonal element of $A_{ik}(t)$ and $B_{ik}(t)$. We break down the second level weight by dynamic grouping based on the sign of $A_{ik}(t)$ and $B_{ik}(t)$. Specifically, denote $\gamma_{ik}^{+}(t) = I(A_{ik}(t) \geq 0)$, $\gamma_{ik}^{-}(t) = I(A_{ik}(t) < 0)$, and let $A_{ik}^{+}(t) = \gamma_{ik}^{+}(t)A_{ik}(t)$, $A_{ik}^{-}(t) = -\gamma_{ik}^{-}(t)A_{ik}(t)$. We then have an estimate of α using only the second level weights that are non-negative:

$$\hat{\alpha}_k^{+}(t) = \{\sum_i (1 - \Delta_{ik})A_{ik}^{+}(t)\}^{-1}\{\sum_i \xi_i(1 - \Delta_{ik})A_{ik}^{+}(t)\}.$$

$\hat{\alpha}_k^{-}(t)$ is defined similarly. For the second level weights $B_{ik}(t)$, we analogously define the quantities:

$$\zeta_{ik}^{+}(t) = I(B_{ik}(t) \geq 0), B_{ik}^{+}(t) = \zeta_{ik}^{+}(t)B_{ik}(t);$$

$$\zeta_{ik}^{-}(t) = I(B_{ik}(t) < 0), B_{ik}^{-}(t) = -\zeta_{ik}^{-}(t)B_{ik}(t);$$

$$\hat{q}_k^{+}(t) = \{\sum_{i=1}^{n} \Delta_{ik}(1 - \xi_i)B_{ik}^{+}(t)\}^{-1}\{\sum_{i=1}^{n} \Delta_{ik}(1 - \xi_i)\eta_{ik}B_{ik}^{+}(t)\},$$

$$\hat{q}_k^{-}(t) = \{\sum_{i=1}^{n} \Delta_{ik}(1 - \xi_i)B_{ik}^{-}(t)\}^{-1}\{\sum_{i=1}^{n} \Delta_{ik}(1 - \xi_i)\eta_{ik}B_{ik}^{-}(t)\}.$$

Finally, the generalized weight function is

$$\begin{aligned}\tilde{w}_{ik}(t) = \Delta_{ik}\xi_i I_p \quad &+ \quad (1 - \Delta_{ik})\xi_i \times [\gamma_{ik}^{+}(t)\hat{\alpha}_k^{+}(t)^{-1} + \gamma_{ik}^{-}(t)\hat{\alpha}_k^{-}(t)^{-1}] \\ &+ \quad \Delta_{ik}(1 - \xi_i)\eta_{ik} \times [\zeta_{ik}^{+}(t)\hat{q}_k^{+}(t)^{-1} + \zeta_{ik}^{-}(t)\hat{q}_k^{-}(t)^{-1}]\end{aligned}$$

The expressions of asymptotic variance also need to be modified to accommodate the grouping:

$$\begin{aligned} V^I(\beta_0) &= var\Big\{ \sum_k (1-\Delta_{1k}) \int_0^\tau \{R_{1k}(\beta_0,t) - \gamma_{1k}^+(t)\mu_k^+(t)^{-1}A_{1k}^+(t)E^+[(1-\Delta_{1k})R_{1k}(\beta_0,t)] \\ &- \gamma_{1k}^-(t)\mu_k^-(t)^{-1}A_{1k}^-(t)E^-[(1-\Delta_{1k})R_{1k}(\beta_0,t)]\}d\Lambda_{0k}(t)\Big\}, \end{aligned}$$

where $\mu_k^+(t) = E[(1-\Delta_{1k})A_{1k}(t)|A_{1k}(t) \geq 0]$ and $E^+[(1-\Delta_{1k})R_{1k}(\beta_0,t)] = E[(1-\Delta_{1k})R_{1k}(\beta_0,t)|A_{1k} \geq 0]$. $\mu_k^-(t)$ and $E^-[(1-\Delta_{1k})R_{1k}(\beta_0,t)]$ are analogously defined. Also,

$$\begin{aligned} V_k^{II}(\beta_0) &= var\Big\{ M_{\bar{z},1k}(\beta_0) - \int_0^\tau \zeta_{1k}^+(t)\theta_k^+(t)^{-1}B_{1k}^+(t)E^+[dM_{\bar{z},1k}(\beta_0)|\Delta_{1k}=1] \\ &- \int_0^\tau \zeta_{1k}^-(t)\theta_k^-(t)^{-1}B_{1k}^-(t)E^-[dM_{\bar{z},1k}(\beta_0)|\Delta_{1k}=1]\Big|\Delta_{1k}=1,\xi_1=0\Big\}. \end{aligned}$$

$\theta_k^+(t)$, $\theta_k^-(t)$, $E^+[dM_{\bar{z},1k}(\beta_0)|\Delta_{1k}=1]$ and $E^-[dM_{\bar{z},1k}(\beta_0)|\Delta_{1k}=1]$ are computed likewise. Due to the grouping, we need to split the sample to estimate the unknown quantities separately in stratum by the sign of the second level weight. Thus in general, a larger sample size is required to achieve satisfactory asymptotic properties.

3.3.3 Generalization to Stratified Sampling Design

Suppose that a cohort of size n can be partitioned into H mutually exclusive strata based on some first-phase covariates. We extend the method to stratified case-cohort studies, whereby sampling is conducted within each stratum with possibly different sampling probabilities. Specifically, let n_h denote the number of subjects in the hth stratum in the full cohort ($h = 1,\ldots,H$) and $n = n_1 + \cdots + n_H$. Then within the hth stratum, we sample $\tilde{n}_h$ subcohort members via simple random sampling with probability being

$\tilde{\alpha}_h = P(\xi_{hi} = 1) = \tilde{n}_h/n_h$. Total subcohort size $\tilde{n} = \tilde{n}_1 + \cdots + \tilde{n}_H$. Subsequently, for the kth disease outcome within the hth stratum, we sample m_{hk} cases outside the subcohort with probability $\tilde{q}_{hk} = m_{hk}/(n_{hk} - \tilde{n}_{hk})$, where n_{hk} and $\tilde{n}_{hk}$ are the numbers of subjects with the kth disease outcome in the hth stratum in the cohort and in the subcohort, respectively. We consider the following model with the stratified sampling design,

$$\lambda_{hik}(t|Z_{hik}(t)) = Y_{hik}(t)\lambda_{0k}(t)e^{\beta_0^T Z_{hik}(t)}. \tag{3.8}$$

We use superscript/subscript 'ST' to denote the stratified version of quantities. The proposed estimator $\hat{\beta}_{DW}^{ST}$ solves the following estimating equation

$$U_{DW}^{ST}(\beta) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{K}\int_0^\tau \tilde{w}_{hik}(t)\{Z_{hik}(t) - \bar{Z}_{k,DW}(\beta,t)\}dN_{hik}(t) = 0, \tag{3.9}$$

where $\tilde{w}_{hik}(t) = \Delta_{hik}\xi_{hi} + (1-\Delta_{hik})\xi_{hi}\hat{\alpha}_{hk}^{-1}(t) + \Delta_{hik}(1-\xi_{hi})\eta_{hik}\hat{q}_{hk}^{-1}(t)$. Estimating equation (3.9) utilizes weights that are estimated within each sampling stratum. The baseline cumulative hazard function $\Lambda_{0k}(t)$ is estimated by a Breslow-Aalen type estimator $\hat{\Lambda}_{0k}^{ST}(\hat{\beta}_{DW}^{ST}, t)$ where

$$\hat{\Lambda}_{0k}^{ST}(\beta,t) = \int_0^t \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\rho_{hjk}(u)dN_{hjk}(u)}{n\sum_{h=1}^{H}\sum_{j=1}^{n_h}\rho_{hjk}(u)Y_{hjk}(u)e^{\beta^T Z_{hjk}(u)}},$$

where $\rho_{hjk}(u) = \Delta_{hik}\xi_{hi} + (1-\Delta_{hik})\xi_{hi}\hat{\alpha}_{hk}^{KC}(u)^{-1} + \Delta_{hik}(1-\xi_{hi})\eta_{hik}\hat{q}_{hk}^{KC}(u)^{-1}$ is the stratified version of weight function used in Kang and Cai (2009).

Using arguments similar to those in section 3.8, the asymptotic properties of $\hat{\beta}_{DW}^{ST}$ can be derived. It can be shown that $\sqrt{n}(\hat{\beta}_{DW}^{ST} - \beta_0)$ converges to a zero-mean normal

distribution with variance function

$$G^{-1}(\beta_0)\left\{\sum_{h=1}^{H} p_h[Q_h(\beta_0)+\frac{1-\tilde{\alpha}_h}{\tilde{\alpha}_h}V_h^I(\beta_0)+(1-\tilde{\alpha}_h)\sum_{k=1}^{K} pr(\Delta_{1k}=1)\frac{1-\tilde{q}_{hk}}{\tilde{q}_{hk}}V_{hk}^{II}(\beta_0)]\right\}G^{-1}(\beta_0),$$

where $p_h = n_h/n$,

$$\begin{aligned}
Q_h(\beta_0) &= E\Big\{\sum_k M_{\tilde{z},h1k}(\beta_0)\Big\}^{\otimes 2},\\
V_h^I(\beta_0) &= var\Big\{\sum_k(1-\Delta_{h1k})\int_0^\tau\{R_{h1k}(\beta_0,t)\\
&\quad-\mu_{hk}(t)^{-1}A_{h1k}(t)E[(1-\Delta_{h1k})R_{h1k}(\beta_0,t)]\}d\Lambda_{0k}(t)\Big\},\ \text{and}\\
V_{hk}^{II}(\beta_0) &= var\Big\{M_{\tilde{z},h1k}(\beta_0)\\
&\quad-\int_0^\tau\theta_{hk}(t)^{-1}B_{h1k}(t)E[dM_{\tilde{z},h1k}(\beta_0)|\Delta_{h1k}=1]\Big|\Delta_{h1k}=1,\xi_{h1}=0\Big\}.
\end{aligned}$$

3.4. Simulation Studies

We performed extensive simulation studies to examine the performance of the proposed doubly-weighted estimator under finite sample setting. Suppose that a case-cohort study was conducted to investigate disease 1 and 2 ($K = 2$). We considered the following set up. There are three covariates of interest: Z_1 and Z_3 were two first-phase covariates where $Z_1 \sim N(0.3, 0.46^2)$ and $Z_3 \sim N(1, 0.5^2)$; Z_2 was the second-phase covariate which was only available for subcohort members and sampled cases. We assumed that Z_2 had a first-phase continuous surrogate $\tilde{Z}_2$ that followed $N(0.5, 0.5^2)$ distribution. We introduced $Z_4 \sim N(0.5, \sigma_4^2)$ to represent the presence of auxiliary covariates. We set $Z_2 = \tilde{Z}_2 + Z_4 + \epsilon$ where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\tilde{Z}_2, Z_4, \epsilon$ were mutually independent. Therefore, $\sigma^2 = \sigma_4^2 + \sigma_\epsilon^2$ controlled the correlation between Z_2 and its first-phase surrogate $\tilde{Z}_2$. Specifically, we had $corr(Z_2, \tilde{Z}_2) = (2\sqrt{0.5^2 + \sigma^2})^{-1}$.

We assumed that the marginal distribution of T_{ik} is exponential with failure rate $\lambda_{0k}e^{\beta_0^T Z_{ik}}$ where β_0 is the true regression parameter vector. Correlated failure time data were generated from the Clayton-Cuzick model (Clayton and Cuzick 1985), in which the joint survival function of $T_i = (T_{i1}, \ldots, T_{iK})^T$ had the form:

$$S(t_{1i}, \ldots, t_{Ki}|Z_{1i}, \ldots, Z_{Ki}) = \left\{ \sum_{k=1}^{K} exp(\frac{\int_0^{t_{ik}} \lambda_{0k}(t)e^{\beta_0' Z_{ik}} dt}{\theta}) - (K-1) \right\}^{-\theta}.$$

The positive parameter θ measured the strength of correlation among $(T_{i1}, \ldots, T_{iK})$. The relationship between θ and Kendall's τ_θ is $\tau_\theta = 1/(2\theta+1)$. The smaller θ was, the larger the Kendall's τ_θ, hence the stronger the correlation. The baseline hazard functions were set to 0.3 for disease 1 and 0.5 for disease 2 ($K = 2$). Right-censoring time $C_i = C_{i1} = C_{i2}$ was generated from uniform distribution on $[0, r]$, hence censoring percentage was controlled by the parameter r.

3.4.1 Traditional Case-cohort Design

We first examined the performance of doubly-weighted estimator under the traditional case-cohort design. We simulated full study cohort samples of size $n = 3000$ and then selected a subcohort of size 300 or 450 ($\tilde{\alpha} = 0.1$ or 0.15) then collected all the cases outside the subcohort. Right-censoring parameter r was selected so that the event rate was roughly 4% and 7% for disease 1 and 2, respectively. Values 0.05, 0.50, 10 were considered for parameter θ, corresponding to Kendall's τ_θ of 0.91, 0.50, 0.05, to represent strong to weak correlation between the two disease outcomes. Lastly, we set $\sigma_4^2 = 0.2$ and $\sigma_\epsilon^2 = 0.06$ so that $corr(Z_2, \tilde{Z}_2) = 0.7$.

In our simulation where $p = 3$, the first-phase covariates $\hat{Z}_{ik,1}$ and $\hat{Z}_{ik,3}$ were their respective observed values. For the subjects in the subcohort and the cases, the second-

phase covariate $\hat{Z}_{ik,2}$ equaled the observed values, while for non-cases outside the subcohort, their $Z_{ik,2}$ was missing and $\hat{Z}_{ik,2}$ equaled the estimated value. We postulated a linear model to estimate the second-phase covariate $Z_{ik,2}$ for non-subcohort controls. Using the fully observed data on subcohort controls and cases, regressing Z_2 on its surrogate $\tilde{Z}_2$ yielded an R^2 around 0.5. If we incorporated the first-phase covariates Z_1, Z_3 and Z_4, the R^2 increased to 0.85. This mimicked the situation that auxiliary information was used to improve the capability predicting missing Z_2. We then obtained $\hat{Z}_2$ for non-subcohort controls and implemented the doubly-weighted estimator $\hat{\beta}_{DW}$. For comparison purpose, we computed estimator II, denoted $\hat{\beta}_{KC}$, in Kang and Cai (2009). The estimator based on the full cohort $\hat{\beta}_F$, which is not feasible in practice with case-cohort designs, was also obtained as a benchmark. Results presented were based on 2000 simulations for each setting.

We considered two sets of values of true regression parameters $\beta_0 = (0.5, 0.0, 0.2)^T$ and $\beta_0 = (0.5, 1.2, 0.2)^T$. Results summarized in Tables 3.1 and 3.2 show that the doubly-weighted estimator was approximately unbiased. As the subcohort size $\tilde{n}$ increased, the average of the estimated standard error got closer to the empirical standard deviation and the 95% confidence interval had satisfactory coverage rate. More importantly, $\hat{\beta}_{DW}$ could be much more efficient than $\hat{\beta}_{KC}$. The efficiency gain was higher for less correlated data. The relative efficiency was smaller when subcohort size increased, but efficiency gain was still noticeable.

3.4.2 Generalized Case-cohort Design

We then examined the performance of doubly-weighted estimator under generalized case-cohort design with non-rare diseases. In practice, it is common to take a 'balanced' sample in which the numbers of cases and controls are roughly the same. Let the proportion of disease k be P_k. By simple algebra, we obtained that $\tilde{q}_k$, the case sampling proportion to achieve the expected case/control ratio R_k for disease k, is independent of full cohort

size n:

$$\tilde{q}_k = \frac{[(1-P_k)R_k - P_k]\tilde{\alpha}}{P_k(1-\tilde{\alpha})}. \tag{3.10}$$

We considered the full cohort size of 4000. We then selected a subcohort of size 400 or 600 ($\tilde{\alpha} = 0.1$ or 0.15). The right-censoring parameter r was set to 0.25 so that the event rate was 19% for disease 1 and 28% for disease 2. Based on (3.10), the corresponding vectors of q_k to achieve roughly 1:1 case/control ratio were $(0.36, 0.18)$ or $(0.58, 0.28)$, respectively.

We set both $A_{ik}(t)$ and $B_{ik}(t)$ to be the same as in (3.6). Results based on 2000 simulations are presented in Tables 3.3 and 3.4. Both estimators were generally unbiased. However, when the subcohort sampling proportion was below 0.1 (results not presented), the standard deviation of $\hat{\beta}_{DW}$ could not be estimated accurately and the efficiency gain was minimal. This phenomenon echoed our discussion in section 3.3.3 that doubly-weighted estimator requires a larger sample size to obtain stable variance estimator. On the other hand, $\hat{\beta}_{KC}$ yielded good standard deviation estimator regardless of $\tilde{\alpha}$. We could see that doubly-weighted estimator is more efficient than $\hat{\beta}_{KC}$, although the magnitude of efficiency gain was not as large compared to the traditional case-cohort design. The correlation between two diseases did not appear to affect the relative efficiency.

3.5. Data Analysis

We applied the proposed procedures to a data set from the Atherosclerosis Risk in Communities (ARIC) study (Ballantyne et al. 2004, 2005). ARIC was a large cohort study which enrolled 15,792 apparently healthy middle-aged men and women from four US communities. A baseline examination was conducted from 1987 to 1989, with 3 more examinations through 1998. Patients were followed up with incident CHD, including CHD-related death, and ischemic incident stroke, a first definite or probable hospitalized stroke

through 1998. It was of interest to examine whether lipoprotein-associated phospholipase, Lp-PLA_2, was associated with increase risk for incident CHD and ischemic stroke. A total of 12,363 subjects comprised the full cohort for this analysis. In order to preserve stored plasma samples and reduce cost, case-cohort studies were implemented, one for CHD and one for stroke. Lp-PLA_2 was measured in plasma from visit 2 (1990 to 1992) in individuals who subsequently developed CHD or stroke (cases) and in a cohort random sample (subcohort). Those who were still alive or disease-free by 12/31/1998 or lost to follow-up were treated as censored. The subcohort was selected using stratified sampling based on gender, race (white versus black), and age group (below versus above 55). Table 3.5 shows the baseline characteristics at visit 2 among different populations.

In this analysis, the two disease outcomes of interest were incident CHD and incident ischemic stroke. A total of 603 CHD cases and 183 ischemic incident stroke cases, along with 777 subcohort subjects were included in the sample. Due to the overlap in two diseases, the total number of assayed sera samples was 1470. The main exposure of interest was tertile group indicators of Lp-PLA_2 (low/moderate/high Lp-PLA_2, reference level being low Lp-PLA_2 group). Other confounders adjusted for were three first-phase stratum covariates, age at visit 2, gender and race, so that our model was comparable to model 1 in Ballantyne et al. (2004). Finally, we assumed that the covariate effects were disease-specific, resulting a total of 10 parameters.

We implemented the proposed doubly-weighted estimator $\hat{\beta}_{DW}$ with second level weight (3.6). To this end, we built a prediction model for the second-phase covariate Lp-PLA_2 (in mg/dL) among non-subcohort controls. The first-phase covariates used in the regression model on Lp-PLA_2 were race, gender, LDL-C, HDL-C and smoking status (never smoked/former smoker/current smoker). We then assigned their tertile group indicators based on the predicted values. For comparison purpose, we calculated $\hat{\beta}_{KC}$, the estimator in Kang and Cai (2009). We used the stratified version of estimating equation (3.9)

and variance estimators to accommodate the stratified sampling nature of ARIC study. The coefficient estimates, standard errors and associated p-values are presented in table 3.6. There was fair agreement between the two methods in terms of point estimates. The findings matched those reported in Ballantyne et al. (2004, 2005). Efficiency-wise, $\hat{\beta}_{DW}$ outperformed $\hat{\beta}_{KC}$: despite a negligible (no more than 6%) increase in standard errors of 3 parameters, $\hat{\beta}_{DW}$ yielded noticeably more efficient results elsewhere. The most noteworthy finding was regarding high Lp-PLA$_2$ group: using doubly-weighted estimator, we had strong evidence that it was significantly associated with elevated incident CHD risk (HR: 1.729, 95% CI: 1.092, 2.736), compared to low Lp-PLA$_2$ group. On the other hand, $\hat{\beta}_{KC}$ deemed the effect insignificant (HR: 1.567, 95% CI: 0.846, 2.903). Other first-phase risk factors that were statistically associated with elevated risks were advancing age (CHD and stroke), white race (CHD) and male (stroke). Based on $\hat{\beta}_{DW}$, we performed a Wald test with 2 degrees of freedom to compare the corresponding coefficients for the Lp-PLA$_2$ group indicators between the two diseases. The p-value for the Wald test was 0.6580, suggesting the Lp-PLA$_2$ effects for the two diseases were not significantly different.

3.6. Concluding Remarks

When implementing the doubly-weighted estimator with second level weight (3.6), we need to build a predicting model for the unobserved second-phase covariates. In both simulation studies and the real data application, we chose to build the model using linear regression. Kernel regression and polynomial regression with carefully calibrated smoothing parameter can be explored, if flexible forms of the covariates are desired. Other choices of second level weights are possible. For example, in a similar fashion to Qi et al. (2005), we can incorporate Nadaraya-Watson kernel estimator in the second level weight.

Throughout this paper, we have assumed a Cox-type marginal proportional hazards model. The additive hazards models, which model risk differences, has often been used as

an alternative to the proportional hazards model. For data arising from multiple case-cohort studies, Kang et al. (2013) proposed a marginal additive hazards model based on a weighted estimating equation approach. They also considered the generalized case-cohort design. To improve efficiency, extending the proposed doubly-weighted approach to marginal additive hazards model will allow us to make full use of first-phase covariate information, thus may merit further investigation.

In the ARIC study, it was possible for a subject to experience both CHD and incident stroke. In many other studies, a subject can be at risk for all types of events from the beginning, but will not be at risk for any other event immediately after the occurrence of one event. For example, it may be of interest to model deaths due to the disease of interest and deaths due to all other causes simultaneously. If this is the case, we need to consider statistical methods from a competing risks perspective (Sorensen and Andersen 2000). Our proposed doubly-weighted approach could be adapted to the competing risks setting.

3.7. Explicit Form of $D_{DW}(\beta)$

Due to the matrix nature of $S^{(0)}_{k,DW}(\beta, t)$, special attention is required to compute the Hessian matrix $D_{DW}(\beta)$. Let $l, l' = 1, \ldots, p$, we can explicitly express $\tilde{w}_{ik}(t)$ and $Z_{ik}(t)$:

$$\tilde{w}_{ik}(t) = diag\{\tilde{w}_{ik,1}(t), \tilde{w}_{ik,2}(t), \ldots, \tilde{w}_{ik,p}(t)\},$$
$$Z_{ik}(t) = [Z_{ik,1}(t), Z_{ik,2}(t), \ldots, Z_{ik,p}(t)]^T.$$

Define the scalar functions

$$S^{(0)}_{k,DW,l}(\beta, t) = n^{-1} \sum_{i=1}^{n} \tilde{w}_{ik,l}(t) Y_{ik}(t) exp\{\beta^T Z_{ik}(t)\},$$

$$S^{(1)}_{k,DW,ll'}(\beta,t) = n^{-1}\sum_{i=1}^{n} \tilde{w}_{ik,l}(t) Z_{ik,l'}(t) Y_{ik}(t) exp\{\beta^T Z_{ik}(t)\},$$

and

$$S^{(2)}_{k,DW,ll'}(\beta,t) = n^{-1}\sum_{i=1}^{n} \tilde{w}_{ik,l}(t) Z_{ik,l}(t) Z_{ik,l'}(t) Y_{ik}(t) exp\{\beta^T Z_{ik}(t)\}.$$

Let $V_{k,DW}(\beta,t)$ be the derivative of $-\bar{Z}_{k,DW}(\beta,t)$ with respect to β. We have

$$D_{DW}(\beta) = \frac{\partial U_{DW}(\beta)}{\partial \beta^T} = \sum_{k=1}^{K} \int_0^{\tau} \tilde{w}_{ik}(t) V_{k,DW}(\beta,t) d\sum_{i=1}^{n} N_{ik}(t),$$

where the lth row of $V_{k,DW}(\beta,t)$ has the form

$$S^{(0)}_{k,DW,l}(\beta,t)^{-2}\Big\{ S^{(1)}_{k,DW,ll}(\beta,t)[S^{(1)}_{k,DW,l1}(\beta,t),\ldots,S^{(1)}_{k,DW,lp}(\beta,t)]$$
$$-[S^{(2)}_{k,DW,l1}(\beta,t),\ldots,S^{(2)}_{k,DW,lp}(\beta,t)]S^{(0)}_{k,DW,l}(\beta,t)\Big\}.$$

3.8. Proof of Theorem 3.3.1

The following two lemmas are important in deriving the asymptotic results and are applied repeatedly.

Lemma 3.8.1. *Let $\xi = (\xi_1,...,\xi_n)^T$ be a random vector containing $\tilde{n}$ ones and $n-\tilde{n}$ zeros, with each permutation equally likely. Let $B_i(t), i = 1,...,n$ be independent and identically distributed real-valued random processes on $[0,\tau]$ with $E[B_i(t)] = \mu_B(t)$, $var(B_i(0)) < \infty$ and $var(B_i(\tau)) < \infty$. Let $B(t) = \{B_1(t),...,B_n(t)\}^T$ and ξ be independent. Suppose that almost all paths of $B_i(t)$ have finite variation. Then, $n^{-1/2}\sum_{i=1}^{n}\xi_i\{B_i(t)-\mu_B(t)\}$ converges weakly in $l^{\infty}[0,\tau]$ to a zero-mean Gaussian process and therefore*

$$n^{-1}\sum_{i=1}^{n}\xi_i\{B_i(t)-\mu_B(t)\} \xrightarrow{p} 0$$

uniformly in t.

This lemma is stated in Lemma A1 in Kang and Cai (2009). Its proof involves the central limit theorem for finite population sampling from Hájek (1960) and example 3.6.14 of van der Vaart (1996). A special case of this lemma is obtained by setting $\xi = J_n$ where J_n is an n-vector of ones.

We need the following results on the asymptotic properties of $\hat{\alpha}_k(t)$ and $\hat{q}_k(t)$. We hereafter present and prove Lemma B3.8.2, Lemma B3.8.4 and Theorem 3.3.1 assuming a single covariate in (3.1). With multiple covariates, $\hat{\alpha}_k(t)$, $\hat{q}_k(t)$ and $S^{(0)}_{k,DW}(\beta, t)$ are p-by-p diagonal matrices, and the arguments below pertain to each of the p processes on the diagonal.

Lemma 3.8.2.

$$n^{1/2}(\hat{\alpha}_k(t)^{-1} - \tilde{\alpha}^{-1}) = \{\tilde{\alpha}\mu_k(t)\}^{-1} n^{-1/2} \sum_{i=1}^{n} (1 - \xi_i/\tilde{\alpha})(1 - \Delta_{ik}) A_{ik}(t) + o_p(1), \quad (3.11)$$

in which $\mu_k(t)$ *is defined as* $E[(1 - \Delta_{1k})A_{1k}(t)]$. *Also, we have similar results for* $\hat{q}_k(t)$*:*

$$n^{1/2}(\hat{q}_k(t)^{-1} - \tilde{q}_k^{-1}) = \{\tilde{q}_k(1-\tilde{\alpha})\theta_k(t)\}^{-1} n^{-1/2} \sum_{i=1}^{n} (1 - \eta_{ik}/\tilde{q}_k)\Delta_{ik}(1-\xi_i)B_{ik}(t) + o_p(1), \quad (3.12)$$

in which $\theta_k(t) = E[\Delta_{1k} B_{1k}(t)]$.

The detailed proof for equation (3.11), which utilizes assumptions 3.3.10 to 3.3.12, the special case of Lemma B3.8.1 and functional delta method, can be found in Kulich and Lin (2004). Since we still have identical and independently cases within each disease k, $k = 1, \ldots, K$, the proof can go through without modification. Equation (3.12) can be shown analogously.

We need another technical lemma from Lin (2000b).

Lemma 3.8.3. *Let $W(t)$ and $Z(t)$ be two sequences of bounded processes. Suppose that $W(t)$ is monotone and converges to $w(t)$ uniformly in t in probability and that $Z(t)$ converges weakly to a zero-mean process with continuous sample paths. Then*

$$\int_0^t \{W(u) - w(u)\} dZ(u) \to 0, \quad \int_0^t Z(u) d\{W(u) - w(u)\} \to 0$$

uniformly in t in probability.

The next lemma states the uniform convergence of $\bar{Z}_{k,DW}(\beta, t)$, to the limit of its full cohort counterpart.

Lemma 3.8.4. *(Convergence of the at-risk average process) For any k,*

$$\sup_{\beta,t} \left\| \bar{Z}_{k,DW}(\beta, t) - \bar{z}_k(\beta, t) \right\| \to_p 0.$$

Proof. We first show that $\sup_{\beta,t} \|S^{(d)}_{k,DW}(\beta, t) - S^{(d)}_{k,F}(\beta, t)\| \to_p 0$ uniformly in t and β for $d = 0, 1$. We start with

$$S^{(d)}_{k,DW}(\beta, t) - S^{(d)}_{k,F}(\beta, t) = n^{-1} \sum_i \{\tilde{w}_{ik}(t) - 1\} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t)$$

Expand the weight function $\tilde{w}_{ik}(t)$ and rearrange terms on the right-hand side (RHS), we

get

$$
\begin{aligned}
S^{(d)}_{k,DW}(\beta,t) - S^{(d)}_{k,F}(\beta,t) &= n^{-1}\sum_i (\frac{\xi_i}{\tilde{\alpha}} - 1) Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \\
&\quad - n^{-1}\sum_i (\frac{\eta_{ik}}{\tilde{q}_k} - 1)\Delta_{ik}\xi_i Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \\
&\quad - n^{-1}\sum_i (\frac{\xi_i}{\tilde{\alpha}} - 1)\Delta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \\
&\quad + n^{-1}\sum_i (\frac{\eta_{ik}}{\tilde{q}_k} - 1)\Delta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \\
&\quad + n^{-1}\sum_i (\hat{\alpha}_k(t)^{-1} - \tilde{\alpha}^{-1})(1-\Delta_{ik})\xi_i Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \\
&\quad + n^{-1}\sum_i (\hat{q}_k(t)^{-1} - \tilde{q}_k^{-1})\Delta_{ik}(1-\xi_i)\eta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t).
\end{aligned}
$$

Taking the norm on both sides,

$$
\begin{aligned}
&\left\| S^{(d)}_{k,DW}(\beta,t) - S^{(d)}_{k,F}(\beta,t) \right\| \\
\leq &\left\| n^{-1}\sum_i (\frac{\xi_i}{\tilde{\alpha}} - 1) Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\| && (3.13) \\
+ &\left\| n^{-1}\sum_i (\frac{\eta_{ik}}{\tilde{q}_k} - 1)\Delta_{ik}\xi_i Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\| && (3.14) \\
+ &\left\| n^{-1}\sum_i (\frac{\xi_i}{\tilde{\alpha}} - 1)\Delta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\| && (3.15) \\
+ &\left\| n^{-1}\sum_i (\frac{\eta_{ik}}{\tilde{q}_k} - 1)\Delta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\| && (3.16) \\
+ &\left\| n^{-1}\sum_i (\hat{\alpha}_k(t)^{-1} - \tilde{\alpha}^{-1})(1-\Delta_{ik})\xi_i Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\| && (3.17) \\
+ &\left\| n^{-1}\sum_i (\hat{q}_k(t)^{-1} - \tilde{q}_k^{-1})\Delta_{ik}(1-\xi_i)\eta_{ik} Z^d_{ik}(t) e^{\beta^T Z_{ik}(t)} Y_{ik}(t) \right\|. && (3.18)
\end{aligned}
$$

We now show each of the six terms converges to 0 in probability uniformly in β and t. (3.13) converges to 0 in probability uniformly in t by the special case of Lemma B3.8.1.

Specifically,

$$\begin{aligned}&\|n^{-1}\sum_i(\frac{\xi_i}{\tilde{\alpha}}-1)Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)\|\\=&\|n^{-1}\sum_i\frac{\xi_i}{\tilde{\alpha}}Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)-n^{-1}\sum_i Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)\|.\end{aligned}$$

By iterated expectation argument conditioning on everything but ξ_i, it is clear that

$$E[\frac{\xi_i}{\tilde{\alpha}}Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)]=E[Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)]=\mu_B(t).$$

By assumption, $\frac{\xi_i}{\tilde{\alpha}}Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)$ has finite variation on $[0,\tau]$. Hence, the aforementioned lemma guarantees the convergence of (3.13) to 0, uniformly in t and β. Through similar arguments, (3.14) - (3.16) converges to 0 in probability uniformly in t and β, respectively.

We then show that (3.17) converges to 0 in probability uniformly in t and β. By Cauchy-Schwarz inequality,

$$\begin{aligned}&\|n^{-1}\sum_i(\hat{\alpha}_k(t)^{-1}-\tilde{\alpha}^{-1})(1-\Delta_{ik})\xi_i Z_{ik}^d(t)e^{\beta^T Z_{ik}(t)}Y_{ik}(t)\|\\\leq&\|\hat{\alpha}_k(t)^{-1}-\tilde{\alpha}^{-1}\|\cdot n^{-1}\sum_i(1-\Delta_{ik})\xi_i\|Z_{ik}^d(t)\|e^{\beta^T Z_{ik}(t)}Y_{ik}(t),\end{aligned}$$

The latter converges to 0 in probability, uniformly in t and β. This can be justified by noting $\hat{\alpha}_k(t)^{-1}-\tilde{\alpha}^{-1}$ converges to 0 in probability uniformly in t, in view of Lemma B3.8.2. Also by Lindeberg condition 3.3.8, $n^{-1}\sum_i(1-\Delta_{ik})\xi_i\|Z_{ik}^d(t)\|e^{\beta^T Z_{ik}(t)}Y_{ik}(t)$ converges to a finite quantity. Likewise, (3.18) can be shown to converge to 0 in probability uniformly in t and β. Therefore, we have shown that $\sup_{\beta,t}\|S_{k,DW}^{(d)}(\beta,t)-S_{k,F}^{(d)}(\beta,t)\|\to_p 0$ uniformly

in t and β, for $d = 0, 1$. This result, in combination with assumption 3.3.6, lead to the conclusion that $\sup_{\beta,t} \|S^{(d)}_{k,DW}(\beta, t) - s^{(d)}_k(\beta, t)\| \to_p 0$ uniformly in t and β.

To obtain the main result of the lemma, we start with

$$\begin{aligned}\sup_{\beta,t} \left\|\bar{Z}_{k,DW}(\beta, t) - \bar{z}_k(\beta, t)\right\| &= \sup_{\beta,t} \left\|\bar{Z}_{k,DW}(\beta, t) - \bar{Z}_{k,F}(\beta, t) + \bar{Z}_{k,F}(\beta, t) - \bar{z}_k(\beta, t)\right\| \\ &\leq \sup_{\beta,t} \left\|\bar{Z}_{k,DW}(\beta, t) - \bar{Z}_{k,F}(\beta, t)\right\| \\ &+ \sup_{\beta,t} \left\|\bar{Z}_{k,F}(\beta, t) - \bar{z}_k(\beta, t)\right\|\end{aligned}$$

Clearly, the second term on the right RHS of the inequality converges to 0 in probability based on full data results. The first term can be written as:

$$\begin{aligned}&\sup_{\beta,t} \left\| \frac{S^{(0)}_{k,F}(\beta, t)\{S^{(1)}_{k,DW}(\beta, t) - S^{(1)}_{k,F}(\beta, t)\} + S^{(1)}_{k,F}(\beta, t)\{S^{(0)}_{k,F}(\beta, t) - S^{(0)}_{k,DW}(\beta, t)\}}{S^{(0)}_{k,DW}(\beta, t)S^{(0)}_{k,F}(\beta, t)} \right\| \\ \leq &\sup_{\beta,t} \left\| \frac{S^{(0)}_{k,F}(\beta, t)\{S^{(1)}_{k,DW}(\beta, t) - S^{(1)}_{k,F}(\beta, t)\}}{S^{(0)}_{k,DW}(\beta, t)S^{(0)}_{k,F}(\beta, t)} \right\| \\ + &\sup_{\beta,t} \left\| \frac{S^{(1)}_{k,F}(\beta, t)\{S^{(0)}_{k,F}(\beta, t) - S^{(0)}_{k,DW}(\beta, t)\}}{S^{(0)}_{k,DW}(\beta, t)S^{(0)}_{k,F}(\beta, t)} \right\|.\end{aligned}$$

Both terms converge to 0 in probability by assumption 3.3.6 and that $\sup_{\beta,t} \|S^{(d)}_{k,DW}(\beta, t) - S^{(d)}_{k,F}(\beta, t)\| \to_p 0$ uniformly in t and β, for $d = 0, 1$. This completes the proof. □

We are now in place of proving theorem 3.3.1.

Proof. The proof of consistency of $\hat{\beta}_{DW}$ can be shown by the extension of Foutz (1977). Denote $n^{-1}U_{DW}(\beta)$ by $\tilde{U}_{DW}(\beta)$. $\hat{\beta}_{DW}$ is consistent if all four conditions below hold: (i) $\partial\tilde{U}_{DW}(\beta)/\partial\beta^T$ exists and is continuous in an open neighborhood $\mathcal{B}$ of β_0; (ii) $\partial\tilde{U}_{DW}(\beta)/\partial\beta^T$ is negative definite with probability going to one as $n \to \infty$; (iii) $-\partial\tilde{U}_{DW}(\beta)/\partial\beta^T$ converges to $G(\beta_0)$ in probability uniformly for β in an open neighborhood of β_0; (iv) $\tilde{U}_{DW}(\beta)$

converges to 0 in probability.

We need to verify the four conditions to establish consistency. The form of $\partial\tilde{U}_{DW}(\beta)/\partial\beta^T$ was given in section 3.7, hence (i) holds due to the continuity of each part. (ii) and (iii) are satisfied if we can show $\| - \partial\tilde{U}_{DW}(\beta)/\partial\beta^T - G(\beta)\|$ converges to 0 in probability uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. We make the decomposition

$$\begin{aligned}\left\| -\frac{\partial\tilde{U}_{DW}(\beta)}{\partial\beta^T} - G(\beta) \right\| &\leq \left\| \sum_{k=1}^{K} \int_0^\tau \{V_{k,DW}(\beta,t) - v_k(\beta,t)\} n^{-1} d \sum_{i=1}^{n} N_{ik}(t) \right\| \\ &+ \left\| \sum_{k=1}^{K} \int_0^\tau v_k(\beta,t) n^{-1} d \sum_{i=1}^{n} M_{ik}(t) \right\| \\ &+ \left\| \sum_{k=1}^{K} \int_0^\tau v_k(\beta,t) \{S^{(0)}_{k,DW}(\beta,t) - s^{(0)}_k(\beta,t)\} d\Lambda_{0k}(t) \right\| \end{aligned} \quad (3.19)$$

Each term on the RHS of (3.19) will be shown to converge to 0, uniformly in $\beta \in \mathcal{B}$. While proving Lemma B3.8.4, we showed that $\sup_{\beta,t} \|S^{(d)}_{k,DW}(\beta,t) - s^{(d)}_k(\beta,t)\| \to_p 0$ uniformly in t and β, for $d = 0, 1$. From the derivation in section 3.7, it follows naturally that $V_{k,DW}(\beta,t)$ converges to $v_k(\beta,t)$ uniformly in t and β. By Lenglart inequality, for any $\delta, \rho > 0$, there exists n_0 such that for $n \geq n_0$,

$$P[n^{-1}\bar{N}_k(\tau) > c] \leq \frac{\delta}{c} + P[\int_0^\tau S^{(0)}_{k,DW}(\beta_0,t)\lambda_{0k}(t)dt > \delta].$$

By assumption 3.3.6, for $\delta > \int_0^\tau s^{(0)}_k(\beta_0,t)\lambda_{0k}(t)dt$, $P[\int_0^\tau S^{(0)}_{k,DW}(\beta_0,t)\lambda_{0k}(t)dt > \delta] \to 0$ as $n \to \infty$. Then $\lim_{c\uparrow\infty} \lim_{n\to\infty} P[n^{-1}\bar{N}_k(\tau) > c] = 0$. Therefore, the first term on the RHS converges to 0 in probability uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$.

For the second term, $n^{-1}\sum_{i=1}^{n} \int_0^\tau v_k(\beta,t) dM_{ik}(t)$ is a local square integrable martin-

gale. Lenglart inequality implies that, for any $\delta, \rho > 0$, there exists n_0 such that for $n \geq n_0$,

$$\begin{aligned} & P\left[\left\| n^{-1} \int_0^\tau \{v_k(\beta, t)\}_{ll'} d\bar{M}_k(t) \right\| > \rho \right] \\ \leq & \frac{\delta}{\rho^2} + P\left[n^{-1} \int_0^\tau \{v_k(\beta, t)\}_{ll'}^2 S_{k,DW}^{(0)}(\beta_0, t) \lambda_{0k}(t) dt > \delta \right] \end{aligned}$$

where the subscript ll' denotes the (l, l') element of the matrix. Assumptions 3.3.5-3.3.7 ensure that $P[n^{-1} \int_0^\tau \{v_k(\beta, t)\}_{ll'}^2 S_{k,DW}^{(0)}(\beta_0, t) \lambda_{0k}(t) dt > \delta]$ converges to 0 in probability uniformly in $\beta \in \mathcal{B}$ for any δ. Then the second term on the RHS of (3.19) also converges to 0 in probability uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$, since δ can be arbitrarily small.

Finally, assumptions 3.3.4-3.3.6 and uniform convergence of $S_{k,DW}^{(0)}(\beta, t)$ to $s_k^{(0)}(\beta, t)$ in probability, the last term on the RHS of (3.19) converges to 0 uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. Therefore, left-hand side (LHS) of (3.19) converges to 0 uniformly in $\beta \in \mathcal{B}$ as $n \to \infty$. Then conditions (ii) and (iii) are satisfied.

Convergence of $\tilde{U}_{DW}(\beta)$ to zero in probability shows that (iv) is satisfied. Therefore, $\hat{\beta}_{DW}$ is a consistent estimator of β_0.

To establish the asymptotic normality of the doubly-weighted score process, we make

the decomposition of $n^{-1/2}U_{DW}(\beta_0)$

$$
\begin{aligned}
n^{-1/2}U_{DW}(\beta_0) &= n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \tilde{w}_{ik}(t)\{Z_{ik}(t)-\bar{Z}_{k,DW}(\beta,t)\}dN_{ik}(t) \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \tilde{w}_{ik}(t)\{Z_{ik}(t)-\bar{Z}_{k,DW}(\beta,t)\}dM_{ik}(t) \\
&= n^{-1/2}\sum_{k}\sum_{i}\int_0^\tau \{Z_{ik}(t)-\bar{z}_{k}(\beta_0,t)\}dM_{ik}(t) \qquad (3.20) \\
&+ n^{-1/2}\sum_{k}\sum_{i}\int_0^\tau \{\bar{z}_{k}(\beta_0,t)-\bar{Z}_{k,DW}(\beta,t)\}dM_{ik}(t) \qquad (3.21) \\
&+ n^{-1/2}\sum_{k}\sum_{i}\int_0^\tau (\tilde{w}_{ik}(t)-1)\{Z_{ik}(t)-\bar{z}_{k}(\beta_0,t)\}dM_{ik}(t) \qquad (3.22) \\
&+ n^{-1/2}\sum_{k}\sum_{i}\int_0^\tau (\tilde{w}_{ik}(t)-1)\{\bar{z}_{k}(\beta_0,t)-\bar{Z}_{k,DW}(\beta,t)\}dM_{ik}(t) \\
& \qquad (3.23) \\
&+ o_p(1).
\end{aligned}
$$

Using the example in 2.11.16 of van der Vaart (1996), the Kolmogorov-Centsov theorem, Lemma B3.8.2 and B3.8.3, (3.21) and (3.23) can be shown to converge to 0 in probability, uniformly in t. In Spiekerman and Lin (1998), (3.20) was shown to converge to a zero mean normal distribution with covariance matrix $Q(\beta_0)$, where $Q(\beta_0) = E[\sum_{k=1}^{K}\int_0^\tau \tilde{Z}_{ik}(\beta,t)dM_{ik}(t)]^{\otimes 2}$.

We can further decompose (3.22) by expanding $\tilde{w}_{ik}(t)$:

$$
\begin{aligned}
& n^{-1/2}\sum_k\sum_i\int_0^\tau(\tilde{w}_{ik}(t)-1)dM_{\bar{z},ik}(\beta_0,t) \\
= \;& n^{-1/2}\sum_k\sum_i(1-\Delta_{ik})\xi_i\int_0^\tau(\hat{\alpha}_k^{-1}(t)-\tilde{\alpha}^{-1})dM_{\bar{z},ik}(\beta_0,t) && (3.24)\\
+ \;& n^{-1/2}\sum_k\sum_i\Delta_{ik}(1-\xi_i)\eta_{ik}\int_0^\tau(\hat{q}_k^{-1}(t)-\tilde{q}_k^{-1})dM_{\bar{z},ik}(\beta_0,t) && (3.25)\\
+ \;& n^{-1/2}\sum_k\sum_i(1-\Delta_{ik})(\xi_i\tilde{\alpha}^{-1}-1)M_{\bar{z},ik}(\beta_0) && (3.26)\\
+ \;& n^{-1/2}\sum_k\sum_i\Delta_{ik}(1-\xi_i)(\eta_{ik}\tilde{q}_k^{-1}-1)M_{\bar{z},ik}(\beta_0). && (3.27)
\end{aligned}
$$

By (3.11), (3.24) is equal to

$$
\begin{aligned}
& n^{-1/2}\sum_k\sum_i(1-\Delta_{ik})\xi_i\times \\
& \int_0^\tau[\{\tilde{\alpha}\mu_k(t)\}^{-1}n^{-1}\sum_j(1-\xi_j\tilde{\alpha}^{-1})(1-\Delta_{jk})A_{jk}(t)]\tilde{Z}_{ik}(\beta_0,t)dM_{ik}(t) \\
= & -n^{-1/2}\sum_k\sum_i(1-\Delta_{ik})(\xi_i\tilde{\alpha}^{-1}-1)\times \\
& \int_0^\tau\mu_k(t)^{-1}A_{ik}(t)\{n^{-1}\sum_j\xi_j\tilde{\alpha}^{-1}(1-\Delta_{jk})\tilde{Z}_{jk}(\beta_0,t)dM_{jk}(t)\} \\
= & n^{-1/2}\sum_k\sum_i(1-\Delta_{ik})(\xi_i\tilde{\alpha}^{-1}-1)\times \\
& \int_0^\tau\mu_k(t)^{-1}A_{ik}(t)\{n^{-1}\sum_j\xi_j\tilde{\alpha}^{-1}(1-\Delta_{jk})\tilde{Z}_{jk}(\beta_0,t)Y_{jk}(t)e^{\beta_0^TZ_{jk}(t)}\}d\Lambda_{0k}(t).
\end{aligned}
$$

The last equation is granted by martingale decomposition of $M_{jk}(t)$ and the fact $(1-$

$\Delta_{jk})dN_{jk}(t) = 0$. Similarly, we have (3.26) equal to

$$\begin{aligned} &- n^{-1/2} \sum_k \sum_i (1 - \Delta_{ik})(\xi_i \tilde{\alpha}^{-1} - 1) \tilde{Z}_{ik}(\beta_0, t) Y_{ik}(t) e^{\beta_0^T Z_{ik}(t)} d\Lambda_{0k}(t) \\ = &- n^{-1/2} \sum_k \sum_i (1 - \Delta_{ik})(\xi_i \tilde{\alpha}^{-1} - 1) R_{ik}(\beta_0, t) d\Lambda_{0k}(t). \end{aligned}$$

The quantity $n^{-1} \sum_j \xi_j \tilde{\alpha}^{-1}(1 - \Delta_{jk}) \tilde{Z}_{jk}(\beta_0, t) Y_{jk}(t) e^{\beta_0^T Z_{jk}(t)}$ converge in probability to $E[(1 - \Delta_{1k}) R_{1k}(\beta_0, t)]$ uniformly in t, by the special case of Lemma B3.8.1. We can then combine (3.24) and (3.26) and can show that the combined term is asymptotically equivalent to

$$\begin{aligned} n^{-1/2} \sum_k \sum_i (1 - \Delta_{ik})(\xi_i \tilde{\alpha}^{-1} - 1) \int_0^\tau \{ &\mu_k(t)^{-1} A_{ik}(t) \\ &\cdot E[(1 - \Delta_{1k}) R_{1k}(\beta_0, t)] - R_{ik}(\beta_0, t) \} d\Lambda_{0k}(t). \end{aligned} \tag{3.28}$$

Repeating the above procedure to combine (3.25) and (3.27), their summation is asymptotically equivalent to

$$n^{-1/2} \sum_k \sum_i \Delta_{ik}(1 - \xi_i)(\eta_{ik} \tilde{q}_k^{-1} - 1)[M_{\bar{z},ik}(\beta_0) - \int_0^\tau \theta_k(t)^{-1} B_{ik}(t) E[\Delta_{1k} dM_{\bar{z},k1}(\beta_0, t)]]. \tag{3.29}$$

By Lemma B3.8.1 and B3.8.2, both (3.28) and (3.29) can be shown to converge to a zero mean normal distribution, respectively.

By law of total expectation, the three terms (3.20), (3.28) and (3.29) are pairwise uncorrelated, which implies independence under normality. Specifically, the covariances between (3.28) and (3.29), (3.20) and (3.29) are both 0 by conditioning on filtration $\mathcal{F}(\tau)$ and ξ. The covariance between (3.20) and (3.28) is 0 by conditioning on $\mathcal{F}(\tau)$. Therefore,

$n^{-1/2}U_{DW}(\beta_0)$ is asymptotically normally distributed with mean zero and we can compute the contributions of (3.20), (3.28) and (3.29) to the asymptotic variance separately.

Following conditional arguments, the second component (3.28) has asymptotic variance $\frac{1-\tilde{\alpha}}{\tilde{\alpha}}V^I(\beta_0)$, in which $V^I(\beta_0)$ equals

$$var\Big\{\sum_{k=1}^{K}(1-\Delta_{1k})\int_0^{\tau}\{R_{1k}(\beta_0,t)-\mu_k^{-1}(t)A_{1k}(t)E[(1-\Delta_{1k})R_{1k}(\beta_0,t)]\}d\Lambda_{0k}(t)\Big\}.$$

Similarly, the asymptotic variance of (3.29) is $(1-\tilde{\alpha})\sum_{k=1}^{K}pr(\Delta_{1k}=1)\frac{1-\tilde{q}_k}{\tilde{q}_k}V_k^{II}(\beta_0)$, where

$$V_k^{II}(\beta_0)=var\Big\{M_{\bar{z},1k}(\beta_0)-\int_0^{\tau}\theta_k(t)^{-1}B_{1k}(t)E[\Delta_{1k}dM_{\bar{z},1k}(\beta_0,t)]|\Delta_{1k}=1,\xi_1=0\Big\}.$$

The desirable asymptotic distribution of $\hat{\beta}_{DW}$ then follows from the Taylor expansion of $U_{DW}(\hat{\beta}_{DW})$ around β_0 and Slutsky's theorem. □

The quantities $G(\beta_0)$, $Q(\beta_0)$, $\frac{1-\tilde{\alpha}}{\tilde{\alpha}}V^I(\beta_0)$ and $(1-\tilde{\alpha})\sum_{k=1}^{K}pr(\Delta_{1k}=1)V_k^{II}(\beta_0)$ can be consistently estimated by $\hat{G}(\hat{\beta}_{DW})$, $\hat{Q}(\hat{\beta}_{DW})$, $\frac{1-\tilde{\alpha}}{\tilde{\alpha}}\hat{V}^I(\hat{\beta}_{DW})$ and $(1-\tilde{\alpha})\sum_k\hat{pr}(\Delta_{1k}=1)\hat{V}_k^{II}(\hat{\beta}_{DW})$, respectively, where

$$\hat{G}(\beta)=-n^{-1}D_{DW}(\beta),\ \ \hat{Q}(\beta)=n^{-1}\sum_{i=1}^{n}\frac{\xi_i}{\tilde{\alpha}}\Big[\sum_{k=1}^{K}\hat{M}_{\bar{z},ik}(\beta)\Big]^{\otimes 2},$$

where

$$\begin{aligned}\hat{M}_{\bar{z},ik}(\beta) = \Delta_{ik}[Z_{ik}(X_{ik}) - S^{(0)}_{k,DW}(\beta, X_{ik})^{-1} S^{(1)}_{k,DW}(\beta, X_{ik})] \\ - n^{-1} \sum_{j=1}^{n} \frac{\Delta_{jk} Y_{ik}(X_{jk}) e^{\beta^T Z_{ik}(X_{jk})}}{\hat{S}^{(0)}_{k,KC}(\beta, X_{jk})} \rho_{jk}(X_{jk}) \\ \cdot [Z_{ik}(X_{jk}) - S^{(0)}_{k,DW}(\beta, X_{jk})^{-1} S^{(1)}_{k,DW}(\beta, X_{jk})],\end{aligned}$$

$$\begin{aligned}\hat{V}^I(\beta) = n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\tilde{\alpha}} \Bigg[\sum_{k=1}^{K} (1 - \Delta_{ik}) \cdot n^{-1} \sum_{j=1}^{n} \frac{\Delta_{jk}}{\hat{S}^{(0)}_{k,KC}(\beta, X_{jk})} \cdot \rho_{jk}(X_{jk}) \\ \times \{\hat{R}_{ik}(\beta, X_{jk}) - \hat{\mu}_k(X_{jk})^{-1} A_{ik}(X_{jk}) \hat{E}[(1 - \Delta_{1k}) R_{1k}(\beta, X_{jk})]\} \Bigg]^{\otimes 2} \\ - \Bigg[n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\tilde{\alpha}} \sum_{k=1}^{K} (1 - \Delta_{ik}) \cdot n^{-1} \sum_{j=1}^{n} \frac{\Delta_{jk}}{\hat{S}^{(0)}_{k,KC}(\beta, X_{jk})} \cdot \rho_{jk}(X_{jk}) \\ \times \{\hat{R}_{ik}(\beta, X_{jk}) - \hat{\mu}_k(X_{jk})^{-1} A_{ik}(X_{jk}) \hat{E}[(1 - \Delta_{1k}) R_{1k}(\beta, X_{jk})]\} \Bigg]^{\otimes 2},\end{aligned}$$

$$
\begin{aligned}
\hat{V}_k^{II}(\beta) = (n_k - \tilde{n}_k)^{-1} \sum_{i=1}^{n} \frac{\eta_{ik}}{\tilde{q}_k} \Bigg[& \hat{M}_{\bar{z},ik}(\beta) \\
& - (n_k - \tilde{n}_k)^{-1} \sum_{j=1}^{n} \Delta_{jk}(1-\xi_j) \frac{\eta_{jk}}{\tilde{q}_k} \hat{\theta}_k(X_{jk})^{-1} \\
& \cdot B_{ik}(X_{jk}) (Z_{jk}(X_{jk}) - S_{k,DW}^{(0)}(\beta, X_{jk})^{-1} S_{k,DW}^{(1)}(\beta, X_{jk})) \\
& + n^{-1} \sum_{j=1}^{n} \frac{\Delta_{jk}}{\hat{S}_{k,KC}^{(0)}(\beta, X_{jk})} \rho_{jk}(X_{jk}) \hat{\theta}_k(X_{jk})^{-1} \\
& \cdot B_{ik}(X_{jk}) \hat{E}[R_{1k}(\beta, X_{jk}) | \Delta_{1k} = 1, \xi_i = 0] \Bigg]^{\otimes 2} \\
& - \Bigg[(n_k - \tilde{n}_k)^{-1} \sum_{i=1}^{n} \frac{\eta_{ik}}{\tilde{q}_k} \Bigg(\hat{M}_{\bar{z},ik}(\beta) \\
& - (n_k - \tilde{n}_k)^{-1} \sum_{j=1}^{n} \Delta_{jk}(1-\xi_j) \frac{\eta_{jk}}{\tilde{q}_k} \hat{\theta}_k(X_{jk})^{-1} \\
& \cdot B_{ik}(X_{jk}) (Z_{jk}(X_{jk}) - S_{k,DW}^{(0)}(\beta, X_{jk})^{-1} S_{k,DW}^{(1)}(\beta, X_{jk})) \\
& + n^{-1} \sum_{j=1}^{n} \frac{\Delta_{jk}}{\hat{S}_{k,KC}^{(0)}(\beta, X_{jk})} \rho_{jk}(X_{jk}) \hat{\theta}_k(X_{jk})^{-1} \\
& \cdot B_{ik}(X_{jk}) \hat{E}[R_{1k}(\beta, X_{jk}) | \Delta_{1k} = 1, \xi_i = 0] \Bigg) \Bigg]^{\otimes 2}
\end{aligned}
$$

$$
\hat{R}_{ik}(\beta, t) = \{ Z_{ik}(t) - S_{k,DW}^{(0)}(\beta, t)^{-1} S_{k,DW}^{(1)}(\beta, t) \} Y_{ik}(t) e^{\beta^T Z_{ik}(t)},
$$

$$
\hat{\mu}_k(t) \;\; = \;\; n^{-1} \sum_{i=1}^{n} (1 - \Delta_{ik}) A_{ik}(t),
$$

$$
\hat{E}[(1-\Delta_{1k}) R_{1k}(\beta, t)] \;\; = \;\; n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\tilde{\alpha}} (1 - \Delta_{ik}) \hat{R}_{ik}(\beta, t),
$$

$$\hat{\theta}_k(t) \;=\; n^{-1}\sum_{i=1}^{n}\Delta_{ik}B_{ik}(t),$$

$$\hat{E}[R_{1k}(\beta,t)|\Delta_{1k}=1,\xi_1=0] \;=\; (n_k-\tilde{n}_k)^{-1}\sum_{l=1}^{n}\Delta_{lk}(1-\xi_l)\frac{\eta_{lk}}{\tilde{q}_k}\hat{R}_{lk}(\beta,t).$$

Table 3.1: *Comparison of three estimators: case-cohort design with* $\beta_0 = (0.5, 0.0, 0.2)^T$

			$\hat{\beta}_F$				$\hat{\beta}_{KC}$				$\hat{\beta}_{DW}$				
$\tilde{n}$	τ_θ		Mean	SE	ESD	CR	Mean	SE	ESD	CR	Mean	SE	ESD	CR	$\text{RE}_{DW\|KC}$
300	0.91	β_1	0.500	0.329	0.325	0.95	0.515	0.463	0.450	0.94	0.495	0.329	0.331	0.94	1.98
		β_2	-0.002	0.095	0.095	0.95	-0.003	0.132	0.133	0.94	-0.002	0.104	0.099	0.93	1.61
		β_3	0.199	0.139	0.137	0.94	0.204	0.193	0.190	0.95	0.197	0.138	0.140	0.94	1.96
	0.50	β_1	0.500	0.278	0.277	0.95	0.513	0.421	0.413	0.95	0.495	0.279	0.284	0.94	2.28
		β_2	-0.001	0.083	0.081	0.95	-0.001	0.123	0.122	0.94	-0.001	0.091	0.084	0.93	1.83
		β_3	0.201	0.118	0.117	0.95	0.207	0.178	0.174	0.95	0.199	0.117	0.120	0.95	2.31
	0.05	β_1	0.504	0.257	0.263	0.95	0.517	0.406	0.403	0.95	0.499	0.258	0.270	0.95	2.48
		β_2	-0.002	0.078	0.077	0.95	-0.003	0.121	0.119	0.94	-0.002	0.087	0.080	0.92	1.93
		β_3	0.200	0.110	0.110	0.95	0.206	0.172	0.170	0.95	0.198	0.110	0.114	0.95	2.44
450	0.91	β_1	0.500	0.329	0.325	0.95	0.503	0.419	0.407	0.94	0.497	0.328	0.327	0.94	1.63
		β_2	-0.002	0.095	0.095	0.95	-0.003	0.121	0.120	0.95	-0.002	0.103	0.097	0.93	1.38
		β_3	0.199	0.139	0.137	0.94	0.203	0.174	0.172	0.95	0.197	0.138	0.138	0.94	1.59
	0.50	β_1	0.500	0.278	0.277	0.95	0.500	0.371	0.368	0.95	0.497	0.277	0.280	0.94	1.79
		β_2	-0.001	0.083	0.081	0.95	-0.002	0.110	0.108	0.95	-0.001	0.091	0.083	0.92	1.46
		β_3	0.201	0.118	0.117	0.95	0.205	0.157	0.155	0.95	0.200	0.117	0.118	0.95	1.80
	0.05	β_1	0.504	0.257	0.263	0.95	0.503	0.354	0.357	0.95	0.502	0.257	0.266	0.95	1.90
		β_2	-0.002	0.078	0.077	0.95	-0.003	0.108	0.105	0.94	-0.002	0.086	0.078	0.92	1.58
		β_3	0.200	0.110	0.110	0.95	0.204	0.151	0.150	0.95	0.199	0.110	0.112	0.95	1.88

NOTE: SE, sample standard deviation; ESD, average standard error estimator; CR, estimated standard error coverage rate of the nominal 95% confidence intervals; $\text{RE}_{DW|KC} = SE^2_{KC}/SE^2_{DW}$, efficiency of $\hat{\beta}_{DW}$ relative to $\hat{\beta}_{KC}$. The full cohort contained 3000 subjects.

Table 3.2: *Comparison of three estimators: case-cohort design with* $\beta_0 = (0.5, 1.2, 0.2)^T$

			$\hat{\beta}_F$				$\hat{\beta}_{KC}$				$\hat{\beta}_{DW}$				
$\tilde{n}$	τ_θ		Mean	SE	ESD	CR	Mean	SE	ESD	CR	Mean	SE	ESD	CR	$\text{RE}_{DW\|KC}$
300	0.91	β_1	0.496	0.309	0.311	0.95	0.507	0.515	0.492	0.94	0.482	0.315	0.334	0.95	2.67
		β_2	1.199	0.094	0.093	0.95	1.235	0.167	0.153	0.92	1.223	0.098	0.112	0.95	2.90
		β_3	0.201	0.133	0.131	0.95	0.208	0.216	0.207	0.94	0.195	0.136	0.140	0.95	2.52
	0.50	β_1	0.497	0.267	0.269	0.95	0.507	0.480	0.459	0.93	0.484	0.275	0.290	0.95	3.05
		β_2	1.199	0.084	0.082	0.95	1.235	0.159	0.144	0.91	1.226	0.089	0.101	0.95	3.19
		β_3	0.199	0.114	0.113	0.96	0.210	0.202	0.193	0.94	0.194	0.117	0.122	0.95	2.98
	0.05	β_1	0.501	0.250	0.248	0.95	0.512	0.457	0.442	0.94	0.488	0.258	0.269	0.95	3.14
		β_2	1.201	0.077	0.075	0.94	1.234	0.152	0.137	0.91	1.229	0.083	0.095	0.94	3.35
		β_3	0.201	0.106	0.104	0.95	0.209	0.197	0.186	0.94	0.196	0.110	0.113	0.95	3.21
450	0.91	β_1	0.496	0.309	0.311	0.95	0.504	0.448	0.436	0.95	0.485	0.312	0.324	0.96	2.06
		β_2	1.199	0.094	0.093	0.95	1.222	0.143	0.136	0.93	1.237	0.098	0.104	0.94	2.13
		β_3	0.201	0.133	0.131	0.95	0.207	0.188	0.183	0.94	0.197	0.134	0.136	0.95	1.97
	0.50	β_1	0.497	0.267	0.269	0.95	0.505	0.412	0.400	0.94	0.487	0.270	0.280	0.95	2.33
		β_2	1.199	0.084	0.082	0.95	1.222	0.135	0.126	0.93	1.238	0.088	0.093	0.93	2.35
		β_3	0.199	0.114	0.113	0.96	0.207	0.175	0.168	0.93	0.196	0.115	0.118	0.95	2.32
	0.05	β_1	0.501	0.250	0.248	0.95	0.508	0.391	0.383	0.94	0.491	0.253	0.259	0.95	2.39
		β_2	1.201	0.077	0.075	0.94	1.222	0.128	0.120	0.93	1.241	0.082	0.085	0.93	2.44
		β_3	0.201	0.106	0.104	0.95	0.208	0.170	0.161	0.93	0.197	0.108	0.109	0.95	2.48

NOTE: The full cohort contained 3000 subjects.

Table 3.3: *Comparison of three estimators: generalized case-cohort design with* $\beta_0 = (0.5, 0.0, 0.2)^T$

			$\hat{\beta}_F$				$\hat{\beta}_{KC}$				$\hat{\beta}_{DW}$				
$\tilde{n}$	τ_θ		Mean	SE	ESD	CR	Mean	SE	ESD	CR	Mean	SE	ESD	CR	$\text{RE}_{DW\|KC}$
400	0.91	β_1	0.505	0.149	0.147	0.95	0.525	0.343	0.336	0.96	0.489	0.254	0.270	0.97	1.82
		β_2	0.001	0.043	0.043	0.96	0.000	0.102	0.099	0.94	0.000	0.072	0.081	0.97	2.01
		β_3	0.199	0.059	0.062	0.96	0.201	0.143	0.142	0.96	0.201	0.106	0.114	0.97	1.82
	0.50	β_1	0.508	0.133	0.129	0.95	0.522	0.331	0.323	0.95	0.497	0.247	0.260	0.96	1.80
		β_2	0.000	0.037	0.038	0.96	-0.001	0.096	0.095	0.94	-0.001	0.069	0.077	0.98	1.94
		β_3	0.199	0.051	0.054	0.96	0.199	0.128	0.136	0.96	0.194	0.098	0.110	0.97	1.71
	0.05	β_1	0.513	0.115	0.113	0.95	0.532	0.322	0.311	0.94	0.517	0.233	0.267	0.97	1.91
		β_2	0.001	0.031	0.033	0.96	0.004	0.088	0.091	0.95	0.007	0.066	0.074	0.98	1.78
		β_3	0.198	0.046	0.047	0.96	0.203	0.125	0.131	0.96	0.203	0.100	0.106	0.97	1.56
600	0.91	β_1	0.505	0.149	0.147	0.95	0.515	0.286	0.273	0.94	0.516	0.220	0.219	0.96	1.69
		β_2	0.001	0.043	0.043	0.96	0.001	0.082	0.080	0.94	-0.001	0.058	0.062	0.97	2.00
		β_3	0.199	0.059	0.062	0.96	0.197	0.114	0.115	0.94	0.198	0.086	0.091	0.97	1.76
	0.50	β_1	0.508	0.133	0.129	0.95	0.505	0.272	0.259	0.94	0.504	0.189	0.198	0.97	2.07
		β_2	0.000	0.037	0.038	0.96	0.002	0.076	0.076	0.94	-0.002	0.056	0.059	0.96	1.84
		β_3	0.199	0.051	0.054	0.96	0.199	0.106	0.109	0.95	0.200	0.079	0.084	0.96	1.80
	0.05	β_1	0.513	0.115	0.113	0.95	0.512	0.258	0.246	0.94	0.522	0.178	0.188	0.96	2.10
		β_2	0.001	0.031	0.033	0.96	0.003	0.072	0.072	0.96	0.001	0.049	0.056	0.97	2.16
		β_3	0.198	0.046	0.047	0.96	0.195	0.111	0.104	0.92	0.195	0.075	0.079	0.95	2.19

NOTE: The full cohort contained 4000 subjects.

Table 3.4: *Comparison of three estimators: generalized case-cohort design with* $\beta_0 = (0.5, 1.2, 0.2)^T$

			$\hat{\beta}_F$				$\hat{\beta}_{KC}$				$\hat{\beta}_{DW}$				
$\tilde{n}$	τ_θ		Mean	SE	ESD	CR	Mean	SE	ESD	CR	Mean	SE	ESD	CR	$\text{RE}_{DW\|KC}$
400	0.91	β_1	0.505	0.152	0.146	0.93	0.521	0.347	0.338	0.94	0.501	0.319	0.282	0.92	1.18
		β_2	1.200	0.048	0.046	0.94	1.225	0.100	0.101	0.95	1.183	0.085	0.088	0.94	1.38
		β_3	0.198	0.062	0.061	0.95	0.212	0.141	0.143	0.94	0.199	0.132	0.119	0.91	1.14
	0.50	β_1	0.503	0.134	0.131	0.94	0.534	0.328	0.327	0.95	0.536	0.308	0.272	0.91	1.13
		β_2	1.201	0.041	0.041	0.95	1.224	0.102	0.097	0.94	1.183	0.085	0.086	0.93	1.44
		β_3	0.200	0.055	0.055	0.95	0.202	0.137	0.138	0.95	0.192	0.128	0.113	0.92	1.15
	0.05	β_1	0.502	0.111	0.111	0.95	0.513	0.313	0.309	0.95	0.504	0.291	0.298	0.92	1.16
		β_2	1.200	0.036	0.036	0.95	1.220	0.096	0.091	0.94	1.185	0.084	0.084	0.94	1.31
		β_3	0.201	0.046	0.047	0.95	0.206	0.130	0.130	0.95	0.197	0.128	0.111	0.90	1.03
600	0.91	β_1	0.505	0.152	0.146	0.93	0.511	0.274	0.276	0.95	0.502	0.243	0.216	0.92	1.27
		β_2	1.200	0.048	0.046	0.94	1.209	0.086	0.082	0.93	1.182	0.071	0.071	0.93	1.47
		β_3	0.198	0.062	0.061	0.95	0.197	0.115	0.116	0.95	0.200	0.102	0.091	0.92	1.27
	0.50	β_1	0.503	0.134	0.131	0.94	0.508	0.268	0.264	0.95	0.496	0.235	0.205	0.91	1.30
		β_2	1.201	0.041	0.041	0.95	1.209	0.080	0.079	0.94	1.185	0.068	0.065	0.92	1.38
		β_3	0.200	0.055	0.055	0.95	0.204	0.117	0.111	0.94	0.201	0.095	0.087	0.92	1.52
	0.05	β_1	0.502	0.111	0.111	0.95	0.499	0.241	0.246	0.96	0.499	0.219	0.197	0.91	1.21
		β_2	1.200	0.036	0.036	0.95	1.204	0.076	0.073	0.93	1.184	0.062	0.061	0.94	1.50
		β_3	0.201	0.046	0.047	0.95	0.201	0.103	0.103	0.94	0.201	0.091	0.082	0.92	1.28

NOTE: The full cohort contained 4000 subjects.

Table 3.5: *Baseline Characteristics of ARIC Study*

	CHD (n=604)	Stroke (n = 183)	Subcohort (n = 777)	Full (n = 12,363)
Age (SD), years	58.6 (5.44)	59.7 (5.54)	56.9 (5.57)	56.8 (5.70)
Male Sex, %	67.7	55.7	42.7	42.2
White Race, %	77.1	56.8	75.2	75.6
Lp-PLA_2 (SD), mg/L	0.427 (0.14)	0.451 (0.17)	0.378 (0.13)	N/A
Lp-PLA_2: Moderate †, %	31.5	22.4	33.9	N/A
Lp-PLA_2: High ‡, %	48.0	53.6	34.4	N/A

†Lp-PLA_2 between 0.310 and 0.422 mg/L
‡Lp-PLA_2 above 0.422 mg/L

Table 3.6: *Coefficient Estimates of Disease-Specific Effect Model*

	$\hat{\beta}_{DW}$			$\hat{\beta}_{KC}$		
	Estimate	Std Err	P-value	Estimate	Std Err	P-value
Disease: CHD						
Age in years/10	0.5279	0.1076	$< .0001$	0.4756	0.2020	0.0185
Male	1.0346	0.2411	$< .0001$	0.9798	0.2730	0.0003
White Race	-0.0904	0.2359	0.7016	-0.1692	0.2591	0.5137
Lp-PLA_2: Moderate	0.4135	0.3469	0.2333	0.2573	0.3296	0.4350
Lp-PLA_2: High	0.5474	0.2343	0.0195	0.4490	0.3146	0.1535
Disease: Stroke						
Age in years/10	0.9702	0.2297	$< .0001$	1.0108	0.4175	0.0155
Male	0.6109	0.4134	0.1395	0.4328	0.4413	0.3267
White Race	-0.9571	0.3936	0.0150	-1.2054	0.3906	0.0020
Lp-PLA_2: Moderate	-0.1162	0.6003	0.8465	-0.2028	0.5989	0.7349
Lp-PLA_2: High	0.4435	0.3697	0.2303	0.6961	0.4849	0.1511

CHAPTER 4: MSCM FOR CLUSTERED FAILURE TIMES

4.1. Introduction

It is widely acknowledged that randomized clinical trials are considered the 'gold standard' in assessing the effectiveness of new therapies or drugs. Through randomization, randomized clinical trials are able to balance distributions of subject characteristics across treatment groups, hence remove the confounding. Treatment effect can be estimated simply by comparing outcomes between treated and untreated groups. Despite the popularity, conducting randomized trials is not always ethical, feasible or timely. For example, an active treatment may be considered beneficial to patients, but random assignment may be deemed unethical. As a result, treatment effect is investigated using observational studies. Comparing to randomized trials, the desirable features of observational studies include less restrictive inclusion criteria, possibility of long followup, and relatively low cost (Benson and Hartz 2000).

Recent years have seen increasing interests in observational comparative effective research (CER), mainly due to the growing adoption of electronic medical record (EMR) database. Though the majority of medical records in US hospitals are still paper-based, they are slowly replaced by EMR databases (Jha et al. 2009). EMR data may also come from insurance claims. Such databases are mostly 'big data' and can serve as a rich source of observational data.

However, like any observational study, the use of EMR data to draw causal inference is hindered by its intrinsic confounding. Among all confounding, confounding by indication is considered the most important limitation of observational studies. Confounding

by indication is introduced if prognostic factor(s) can be related to treatment history and outcome. This persistent issue, along with new challenges brought by EMR databases, has drawn great interests from statisticians. Many efforts have been put into the emulation of randomized trials (Danaei et al. 2013). A considerable portion of literature used large observational study databases and applied inclusion/exclusion criteria that mimic the design of well-planned randomized trials. For example, Hernan et al. (2008) compared the results of an observational data (Nurses' Health Study) and a randomized trial (Women's Health Initiative) trial on CHD risk in women from two treatment groups: estrogen/progestin and placebo. Their investigation showed that estimates of the two studies were fairly similar. See also Tannen et al. (2009). Some researchers did literature search or performed meta-analysis to assess the consistency of estimators from observational studies and randomized trials (Ioannidis et al. 2001, Concato et al. 2000).

In randomized clinical trials, the marginal, or unconditional, treatment effect is often 'overwhelmingly the focus of the primary analysis' (Tsiatis et al. 2008). As its name suggests, the marginal treatment effect is not adjusted for other confounders. Loosely speaking, in a two arm randomized trial for example, it can be obtained by fitting a model whose only covariate is the treatment indicator. This type of analysis fits well in the scope of observational CER using EMR databases because, as was pointed out in Sturmer et al. (2011), EMR databases mostly come from hospitals (or insurance companies) and may be lack of data on some important confounders. Some routine risk factors like blood pressure may be available, but other factors that are difficult to measure are likely to be missing.

In many observational studies, whether a subject receives an active treatment or not is determined by a number of individual-level prognostic covariates such as age and co-morbidity. Meanwhile, patients from the same community or clinic form natural clusters. The outcomes of members from the same cluster may be correlated. They may also share a similar tendency to be assigned the active treatment or otherwise. For example, the

INSPIRIS Inc. home visiting provider (HVP) program was initiated to deliver an intensive program that includes home visits by physicians and nurse practitioners and telephonic case management for a high-risk subset of high risk seniors. It is believed that this HVP program has the potential to increase quality of care and reduce total health care expenditures for elders with chronic conditions. Like other studies, individual's medical history and other factors played an important role in determining the program eligibility. Enrollment of HVP program was offered in selected communities in the greater Detroit, Ann Arbor/Lansing and Grand Rapids areas, Michigan. Therefore, subjects living in vicinity form clusters and are potentially correlated. The program was offered to 1,082 participants and claim data are also available on 10,712 non-participants. The investigators are interested in whether the program can improve the quality of life and reduce health insurance claim payments.

Marginal structural models (Robins et al. 2000, Hernan et al. 2001) are a class of models used in causal inference. Such models handle the issue of confounding in evaluation of the efficacy of interventions by inverse probability weighting for receipt of treatment, creating a pseudo-population in which the distribution of prognostic covariates are balanced. Marginal structural Cox model has been employed widely in observational studies, e.g. Hernan et al. (2000). Theoretical justifications of marginal structural models were provided under the assumption of independent observations with continuous responses (Robins 1999) or survival endpoints (Lee 2013). However, when the observations are not independent, these methods do not apply directly. Development of multivariate marginal structural Cox model is needed in order to handle such data properly.

In this paper, we study marginal structural Cox model in the presence of cluster-level random effect. In section 4.2, we formulate the marginal structural Cox model under counter-factual framework. Section 4.3 describes estimation procedure and provides theoretical justification. We carry out an extensive simulation study in section 4.4 and applied the method to a real data set in section 4.5. We conclude with some remarks in section 4.6.

4.2. Statistical Framework

Consider an observational study where the outcome of interest is survival time T. Let $A(t)$ indicate the observed treatment(s) received which can take various forms. For example, it could be an indicator of treatment initiated at baseline, an arbitrary function of dose level, or a time-dependent indicator of treatment received at time t. Let $L(t)$ denote a vector of covariates and $L(0)$ represents baseline covariates. We use overbars to represent history up to time t (t included) such that $\overline{A}(t) = \{A(u) : 0 \leq u \leq t\}$. $\overline{L}(t)$ is defined analogously.

We use the counter-factual outcome framework to formally define the parameters of interest. Let $\overline{a}$ be any treatment, potentially contrary to what was observed, that a subject could receive. Specifically, $\overline{a} = \{a(t) : 0 \leq t \leq \tau\}$, where τ is the duration of the study. Observed treatment history $\overline{A}(t)$ can be considered a particular realization of $\overline{a}(t)$. There is a failure time $T_{\overline{a}}$ associated with each possible realization of $\overline{a}$. The simplest case is when we only consider a binary treatment assignment at baseline. The counter-factual is thus two-dimensional, with two possible outcomes T_1 and T_0, representing the failure time had the subject been assigned to experimental and control group, respectively.

We need the following three assumptions for marginal structural models:

Assumption 4.2.1. $T = T_{\overline{a}}$ *for any* $\overline{a}$ *such that* $a(t) = A(t), t \leq T$.

Assumption 4.2.2. $pr(A(t)|\overline{A}(t^-), \overline{L}(t^-)) > 0$, *for any* $t \in [0, \tau]$ *such that*

$$pr(\overline{A}(t^-), \overline{L}(t^-)) > 0.$$

Assumption 4.2.3. $T_{\overline{a}} \perp\!\!\!\perp A(t)|\overline{A}(t^-), \overline{L}(t^-)$, *for any* $\overline{a}$.

Assumptions 4.2.1 and 4.2.2 are usually referred to as consistency and positivity as-

sumptions, respectively (Hernan and Robins 2006, Cole and Frangakis 2009). Assumption 4.2.1 states that an individual's observed failure time T is precisely the potential failure time $T_{\overline{a}}$ under a certain observed exposure history $\overline{a}$. Assumption 4.2.2 states that the probability of receiving any particular treatment at time t, given treatment and covariate history up to t, is greater than zero. Assumption 4.2.3 is known as no unmeasured confounding (Hernan et al. 2000). In practice, only assumption 4.2.2 is empirically testable.

We consider the marginal structural Cox model for the hazard of failure at time t had the subject received treatment $\overline{a}$

$$\lambda_{T_{\overline{a}}}(t) = \lambda_0(t) exp\{\beta_0^T a(t)\}, \tag{4.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and β_0 is the unknown parameter vector. β_0 will have the interpretation of average treatment log-hazard ratio.

4.3. Estimation and Inference

To facilitate understanding, we hereafter formulate the estimating procedure using the setting in which the patients of the same doctor form a cluster. We use $k = 1, \ldots, K$ to index the doctors. For notational simplicity, we assume that all doctors have n_0 patients. The total sample size $n = K \cdot n_0$. To accommodate the practical situation that numbers of patients vary among doctors, we introduce the indicator $\xi_{ki}, i = 1, \ldots n_0$ which equals 1 if doctor k has patient i, and 0 otherwise. Let T_{ki} and C_{ki} be the potential failure time and censoring time, respectively. Observed time $X_{ki} = T_{ki} \wedge C_{ki}$. The event indicator $\Delta_{ki} = I(T_{ki} < C_{ki})$. Let $Y_{ki}(s) = I(X_{ki} \geq s)$ be the left-continuous at-risk process.

In this paper, we study the estimation and large sample theory under the counting process framework. Let $N_{ki}(t)$ be the counting process representing the number of failures of

subject i in cluster k by time t. We use $dN_{ki}(t)$ to denote the number of events of subject i in cluster k that occurred in $[t, t+dt)$ for some sufficiently small dt. We define the filtration

$$\mathcal{F}_t = \sigma\{N_{ki}(u), Y_{ki}(u)^+, A_{ki}(u), L_{ki}(u); k = 1, \ldots, K, i = 1 \ldots, n_0, 0 \leq u \leq t\},$$

which is the increasing right-continuous σ-algebra generated by failure times, covariates and treatment histories up to time t, and censoring histories up to time t^+ for all subjects. Let $C_{ki}(t) = 0$ indicate that subject i in cluster k remained uncensored prior to time t and $C_{ki}(t) = 1$ otherwise. The treatment process $A_{ki}(\cdot)$ and the censoring process $C_{ki}(\cdot)$ are assumed to be piece-wise constant point processes with cadlag (right-continuous with left-hand limits) step-function sample paths. They are assumed to have jumps that can occur at no more than a finite number of time points. Informally, this means that all participants follow (approximately) the same visit schedule. This assumption should be reasonable in studies with regularly scheduled follow-up visits and good study compliance.

Parameters in model (4.1) are estimated using inverse probability weighting technique. Let $0 \leq t_1 < t_2 < \cdots < t_D \leq \tau$ be D distinct time points, which can be distinct observed times (event or not), or time of scheduled follow-up visits. Define the weight function

$$W_{ki}(t) = \prod_{t_d \leq t} \frac{1}{pr[A_{ki}(t_d)|\overline{A}_{ki}(t_d^-), \overline{L}_{ki}(t_d)]}. \tag{4.2}$$

At any given time t, the subject is inversely weighted by the probability of receiving the observed treatment $A(t)$ conditional on the covariate and treatment histories up to that moment t. By inverse probability weighting, we create a hypothetical pseudo-population where $\overline{L}(t) \perp\!\!\!\perp A(t)|\overline{A}(t)$ holds at time t.

In practice, the true weight function $W_{ki}(t)$ is almost always unknown and needs to be estimated by $\hat{W}_{ki}(t)$. With clustered data, the subjects in the same cluster may share the

same tendency of receiving treatment, we can use methods such as mixed effect models to account for this correlation in estimating $\hat{W}_{ki}(t)$.

Due to the correlated nature of data, the likelihood function that pertains to the original marginal structural model is not applicable to our setting. We propose a modified weighted pseudo partial log-likelihood (WPPL) function

$$l(\beta, t, W) = \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s) \Bigg[\beta^T A_{ki}(s) - log\Bigg\{ n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_0} \xi_{ki} W_{ki}(t) Y_{ki}(t) exp\{\beta^T A_{ki}(t)\} \Bigg\} \Bigg] dN_{ki}(s). \tag{4.3}$$

We substitute $W_{ki}(t)$ by $\hat{W}_{ki}(t)$ in (4.3) to obtain $l(\beta, \tau, \hat{W})$. The maximum WPPL estimator of β_0, denote by $\hat{\beta}_W$, maximizes $l(\beta, \tau, \hat{W})$. $\hat{\beta}_W$ is found via Newton-Raphson algorithm.

When $L(t)$ contains confounders that are strongly correlated to treatment $A(t)$, the estimated weight $\hat{W}(t)$ can vary drastically, resulting in high sampling variability in $\hat{\beta}_W$. As a remedy, we can use a stabilized weight

$$w_{ki}(t) = \prod_{t_d \leq t} \frac{pr[A_{ki}(t_d) | \overline{A}_{ki}(t_d^-)]}{pr[A_{ki}(t_d) | \overline{A}_{ki}(t_d^-), \overline{L}_{ki}(t_d)]}. \tag{4.4}$$

In (4.4), the excessive contribution of $W(t)$ can be offset by the probability conditional on treatment history solely on the numerator. Substituting weight (4.4) in pseudo-partial likelihood (4.3) will yield estimator $\hat{\beta}_w$.

Before presenting results related to the asymptotic distribution of $\hat{\beta}_W$, we introduce

some additional notations. For $d = 0, 1, 2$ and $c = 1, 2$, let

$$S_{W^c}^{(d)}(\beta, t) = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_0} \xi_{ki} W_{ki}^c(t) Y_{ki}(t) A_{ki}(t)^{\otimes d} exp\{\beta^T A_{ki}(t)\},$$

in which $a^{\otimes 0} = 1, a^{\otimes 1} = a, a^{\otimes 2} = aa^T$. $S_{\hat{W}^c}^{(d)}(\beta, t)$ are defined likewise by replacing $W(\beta, t)$ by $\hat{W}(\beta, t)$. For aforementioned quantities with $c = 1$, we suppress c in the notation. In addition, define the following quantities

$$\bar{Z}_W(\beta, t) = \frac{S_W^{(1)}(\beta, t)}{S_W^{(0)}(\beta, t)}; \quad V_W(\beta, t) = \frac{S_W^{(2)}(\beta, t)}{S_W^{(0)}(\beta, t)} - \bar{Z}_W(\beta, t)^{\otimes 2}.$$

The proof of asymptotic results requires the following regularity conditions that are similar to those in Andersen and Gill (1982) and Sasieni (1993):

Assumption 4.3.1. *(Finite interval)* $\int_0^\tau \lambda_0(s) ds < \infty$.

Assumption 4.3.2. *(Asymptotic stability) For* $d = 0, 1, 2$ *and* $c = 1, 2$*, there exists a neighborhood* $\mathcal{B}$ *of* β_0 *and scalar, vector and matrix functions* $s_{W^c}^{(0)}$*,* $s_{W^c}^{(1)}$ *and* $s_{W^c}^{(2)}$ *defined on* $\mathcal{B} \times [0, \tau]$ *such that*

$$\sup_{t \in [0,\tau], \beta \in \mathcal{B}} \| S_{W^c}^{(d)}(\beta, t) - s_{W^c}^{(d)}(\beta, t) \| \to_p 0, \text{as } K \to \infty$$

Assumption 4.3.3. *(Lindeberg condition) There exists* $\delta > 0$ *such that as* $K \to \infty$*,*

$$K^{-1/2} \sup_{k,i,t} \| A_{ki}(t) \| Y_{ki}(t) I\{\beta_0^T A_{ki}(t) > -\delta \| A_{ki}(t) \|\} \to_p 0$$

Assumption 4.3.4. *(Asymptotic regularity condition) Let* $\mathcal{B}$*,* $s_{W^c}^{(0)}$*,* $s_{W^c}^{(1)}$ *and* $s_{W^c}^{(2)}$ *be as in*

assumption 4.3.2 and define

$$\bar{z}(\beta,t)=\frac{s_W^{(1)}(\beta,t)}{s_W^{(0)}(\beta,t)},\quad v(\beta,t)=\frac{s_W^{(2)}(\beta,t)}{s_W^{(0)}(\beta,t)}-\bar{z}(\beta,t)^{\otimes 2}.$$

For all $\beta\in\mathcal{B}, t\in[0,\tau]$*:*

$$s_{W^c}^{(1)}(\beta,t)=\frac{\partial}{\partial\beta}s_{W^c}^{(0)}(\beta,t),\quad s_{W^c}^{(2)}(\beta,t)=\frac{\partial^2}{\partial\beta^2}s_{W^c}^{(0)}(\beta,t).$$

$s_{W^c}^{(0)}(\beta,t), s_{W^c}^{(1)}(\beta,t)$ *and* $s_{W^c}^{(2)}(\beta,t)$ *are continuous functions of* $\beta\in\mathcal{B}$*, uniformly in* $t\in[0,\tau]$*. Further,* $s_{W^c}^{(0)}(\beta,t), s_{W^c}^{(1)}(\beta,t)$ *and* $s_{W^c}^{(2)}(\beta,t)$ *are bounded on* $\mathcal{B}\times[0,\tau]$*;* $s_{W^c}^{(0)}(\beta,t)$ *is bounded away from zero on* $\mathcal{B}\times[0,\tau]$*, and the matrix*

$$G_W(\beta_0)=\int_0^\tau v(\beta_0,t)s_W^{(0)}(\beta_0,t)\lambda_0(t)dt$$

is positive definite.

We need another assumption concerning the weights $W(t)$ and the corresponding estimators $\hat{W}(t)$.

Assumption 4.3.5. *Let* $\|\cdot\|_\infty$ *be the supremum norm over* k,i,t*,*

A. (Uniform consistency of estimated weights)

$$\|\hat{W}_{ki}(t)-W_{ki}(t)\|_\infty\equiv M_W\to_p 0,$$

B. (Stability of weights) Both $W_{ki}(t)$ *and* $\hat{W}_{ki}(t)$ *are locally bounded, that is, there*

exists constants M_1 and M_2 such that

$$\|W_{ki}(t)\|_\infty \leq M_1 \text{ and } \|\hat{W}_{ki}(t)\|_\infty \leq M_2$$

Note that $\hat{W}(t)$ and $W(t)$ are predictable with respect to the filtration $\mathcal{F}_t$. Under the assumption of finite support of $A(\cdot)$, this is true because they are determined by predictable processes: $A(\cdot), L(\cdot)$ and their histories.

Theorem 4.3.1. *(Consistency) Under assumptions 4.3.1 - 4.3.5, $\hat{\beta}_W \to_p \beta_0$, as $K \to \infty$.*

Proof. Consider the following process

$$\begin{aligned} X(\beta, t, W) &= K^{-1}\{l(\beta, t, W) - l(\beta_0, t, W)\} \\ &= K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s) \left[(\beta - \beta_0)^T A_{ki}(s) - log \frac{S_W^{(0)}(\beta, s)}{S_W^{(0)}(\beta_0, s)} \right] dN_{ki}(s), \end{aligned}$$

and its compensator

$$C(\beta, t, W) = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s) \left[(\beta - \beta_0)^T A_{ki}(s) - log \frac{S_W^{(0)}(\beta, s)}{S_W^{(0)}(\beta_0, s)} \right] \lambda_{ki}(s) ds,$$

where the intensity process is given by $\lambda_{ki}(s) = Y_{ki}(s)\lambda_0(s)exp\{\beta_0^T A_{ki}(s)\}$. In this expression, β and $\lambda_0(s)$ in the intensity process for the observed counting process $N(s)$ are the same with the corresponding quantities in model (4.1). To see this, note that the observed counting process is the same with the counting process in the super-population defined by MSCM (4.1). Therefore, the observed and super-population intensity processes have the same β and $\lambda_0(s)$.

To start with, we show that

$$\left|\{X(\beta,t,\hat{W})-C(\beta,t,\hat{W})\}-\{X(\beta,t,W)-C(\beta,t,W)\}\right|\to_p 0 \tag{4.5}$$

so that we can focus on the asymptotic properties of $X(\beta,t,W)-C(\beta,t,W)$ instead of $X(\beta,t,\hat{W})-C(\beta,t,\hat{W})$. The quantity on the left hand side (LHS) of (4.5) can be expressed as

$$\begin{aligned}\Bigg|&K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\{\hat{W}_{ki}(s)-W_{ki}(s)\}(\beta-\beta_0)^T A_{ki}(s)dM_{ki}(s)\\ &-K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\{\hat{W}_{ki}(s)-W_{ki}(s)\}log\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}dM_{ki}(s)\\ -K^{-1}&\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\hat{W}_{ki}(s)log\left\{\frac{S_{\hat{W}}^{(0)}(\beta,s)}{S_{\hat{W}}^{(0)}(\beta_0,s)}\Big/\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}dM_{ki}(s)\Bigg|\end{aligned} \tag{4.6}$$

$W_{ki}(\cdot)$, $\hat{W}_{ki}(\cdot)$ and $A_{ki}(\cdot)$ are all predictable processes, hence each term in (4.6) is a local square integrable martingale. To establish the equivalence (4.5), it suffices to show that the variation process of each martingale in (4.6) converges to 0 in probability. The variation process of the first martingale

$$\begin{aligned}B_1(\beta,t)&=K^{-2}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\{\hat{W}_{ki}(s)-W_{ki}(s)\}^2(\beta-\beta_0)^T A_{ki}(s)^{\otimes 2}(\beta-\beta_0)\lambda_{ki}(s)ds\\ &\le K^{-1}M_W^2\int_0^t(\beta-\beta_0)^T\left[K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\xi_{ki}Y_{ki}(s)exp\{\beta_0^T A_{ki}(s)\}A_{ki}(s)^{\otimes 2}\right]\\ &\cdot(\beta-\beta_0)\lambda_0(s)ds\\ &=K^{-1}M_W^2\int_0^t(\beta-\beta_0)^T S_W^{(2)}(\beta_0,s)(\beta-\beta_0)\lambda_0(s)ds,\end{aligned}$$

The variation process of the second martingale

$$B_2(\beta,t) = K^{-2}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\{\hat{W}_{ki}(s) - W_{ki}(s)\}^2\left\{log\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}^2\lambda_{ki}(s)ds$$
$$\leq K^{-1}M_W^2\int_0^t\left\{log\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}^2 S_W^{(0)}(\beta_0,s)\lambda_0(s)ds,$$

As $K \to \infty$, both $B_1(\beta,t)$ and $B_2(\beta,t)$ converge to 0 in probability, in view of assumptions 4.3.2, 4.3.4 and 4.3.5.

The variation process of the last martingale in (4.6) is given by

$$B_3(\beta,t) = K^{-2}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}\hat{W}_{ki}^2(s)\left[log\left\{\frac{S_{\hat{W}}^{(0)}(\beta,s)}{S_{\hat{W}}^{(0)}(\beta_0,s)}\Big/\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}\right]^2\lambda_{ki}(s)ds$$
$$= K^{-1}\int_0^t\left[log\left\{\frac{S_{\hat{W}}^{(0)}(\beta,s)}{S_{\hat{W}}^{(0)}(\beta_0,s)}\Big/\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}\right]^2 S_{\hat{W}^2}^{(0)}(\beta_0,s)\lambda_0(s)ds$$
$$\leq K^{-1}\int_0^t\left[\left\|log\left\{\frac{S_{\hat{W}}^{(0)}(\beta,s)}{S_{\hat{W}}^{(0)}(\beta_0,s)}\right\}\right\|_\infty^2 + \left\|log\left\{\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}\right\|_\infty^2\right.$$
$$\left. + 2\frac{S_{\hat{W}}^{(0)}(\beta,s)}{S_{\hat{W}}^{(0)}(\beta_0,s)}\right\}\Big\|_\infty\left\|log\left\{\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right\}\right\|_\infty\Big] S_{\hat{W}^2}^{(0)}(\beta_0,s)\lambda_0(s)ds,$$

which converges to 0 in probability by assumptions 4.3.2, 4.3.4, 4.3.5 and continuous mapping theorem, as $K \to \infty$. Hence (4.5) is proved.

The remaining arguments pertain to the martingale $X(\beta,t,W) - C(\beta,t,W)$. We have

$$X(\beta,t,W) - C(\beta,t,W)$$
$$= K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\int_0^t \xi_{ki}W_{ki}(s)\left[(\beta-\beta_0)^T A_{ki}(s) - log\frac{S_W^{(0)}(\beta,s)}{S_W^{(0)}(\beta_0,s)}\right]dM_{ki}(s).$$

Its variation process $B(\beta, t)$ can be expressed as

$$K^{-1}\int_0^t \Bigg[(\beta-\beta_0)^T S^{(2)}_{W^2}(\beta_0,s)(\beta-\beta_0)$$
$$-2(\beta-\beta_0)^T S^{(1)}_{W^2}(\beta_0,s) log\frac{S^{(0)}_W(\beta,s)}{S^{(0)}_W(\beta_0,s)} + \left\{log\frac{S^{(0)}_W(\beta,s)}{S^{(0)}_W(\beta_0,s)}\right\}^2 S^{(0)}_{W^2}(\beta_0,s)\Bigg]\lambda_0(s)ds.$$

By assumptions 4.3.2 and 4.3.4, $K\cdot B(\beta,t)$ converges to some finite quantity involving $s^{(0)}_W(\beta,t)$ and $s^{(d)}_{W^2}(\beta,t)$, for $d=0,1,2$. Hence $B(\beta,t)$ converges to 0 in probability, as $K\to\infty$. Then by Lenglart's inequality,

$$pr\Bigg[\|X(\beta,t,W)-C(\beta,t,W)\|_\infty > \eta\Bigg] \le \frac{\delta}{\eta^2} + pr[B(\beta,\tau)>\delta].$$

It follows that $X(\beta,t,W)$ and $C(\beta,t,W)$ have the same limit when K goes to infinity. As a result, we can investigate the asymptotic properties of $C(\beta,\tau,W)$ instead. Specifically, we examine the first and second order derivatives of $C(\beta,\tau,W)$:

$$\frac{\partial C(\beta,\tau,W)}{\partial\beta} = \int_0^\tau \{S^{(1)}_W(\beta_0,t) - S^{(0)}_W(\beta_0,t)\cdot\frac{S^{(1)}_W(\beta,t)}{S^{(0)}_W(\beta,t)}\}\lambda_0(t)dt,$$

which equals 0 when evaluated at $\beta=\beta_0$. The minus second order derivative

$$\begin{aligned}
-\frac{\partial^2 C(\beta,\tau,W)}{\partial\beta\partial\beta^T} &= -\frac{\partial}{\partial\beta}\{\int_0^\tau (\bar{Z}_W(\beta_0,t)-\bar{Z}_W(\beta,t))S^{(0)}_W(\beta_0,t)\lambda_0(t)dt\}\\
&= \int_0^\tau V_W(\beta,t)S^{(0)}_W(\beta_0,t)\lambda_0(t)dt\\
&\to_p \int_0^\tau v(\beta,t)s^{(0)}_W(\beta_0,t)\lambda_0(t)dt. \qquad (4.7)
\end{aligned}$$

The limiting matrix (4.7) is positive definite, due to assumption 4.3.4.

Therefore, we have shown that $X(\beta, \tau, W)$ converges to a concave function having unique maximum at β_0. By corollary II.2 in Andersen and Gill (1982), we conclude that $\hat{\beta}_W \to_p \beta_0$, as $K \to \infty$. □

Theorem 4.3.2. *(Asymptotic Normality) Under assumptions 4.3.1 - 4.3.5,*

$$K^{1/2}(\hat{\beta}_W - \beta_0) \to_d N(0, G_W(\beta_0)^{-1} G_U(\beta_0) G_W(\beta_0)^{-1}),$$

where

$$G_U(\beta_0) = var\{K^{-1/2} U(\beta_0, \tau, W)\},$$
$$G_W(\beta_0) = \int_0^\tau v(\beta_0, t) s_W^{(0)}(\beta_0, t) \lambda_0(t) dt$$

We obtain the score function by differentiating (4.3) with respect to β:

$$U(\beta, t, W) = \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s)[A_{ki}(s) - \bar{Z}_W^{(0)}(\beta, s)] dN_{ki}(s). \tag{4.8}$$

Differentiate (4.8) with respect to β to obtain information matrix

$$\mathcal{I}(\beta, t) = -\frac{\partial U(\beta, t, W)}{\partial \beta^T} = \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s) V_W(\beta, s) dN_{ki}(s)$$

Expand $U(\hat{\beta}_W, \tau, W)$ in a Taylor series around β_0 to get

$$K^{1/2}(\hat{\beta}_W - \beta_0) = \{K^{-1} \mathcal{I}(\beta_W^*)\}^{-1} K^{-1/2} U(\beta_0, \tau, W), \tag{4.9}$$

where β_W^* is on the line segment between β_0 and $\hat{\beta}_W$.

Lemma 4.3.1. *(Normality of score function) Under assumptions 4.3.1 - 4.3.5, the score*

function $K^{-1/2}U(\beta_0, \tau, W)$ *converges in distribution to a zero mean Gaussian distribution with covariance matrix*

$$G_U(\beta_0) = var\Big\{ \sum_{i=1}^{n_0} \xi_{1i} \int_0^\tau W_{1i}(s)\{A_{1i}(s) - \bar{z}(\beta_0, s)\} dM_{1i}(s) \Big\}$$

Proof. We rewrite (4.8) into the summation of a martingale process and its compensator:

$$\begin{aligned} U(\beta, t, W) = &\sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s)\{A_{ki}(s) - \bar{Z}_W(\beta, s)\} dM_{ki}(s) \\ &+ \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^t \xi_{ki} W_{ki}(s)\{A_{ki}(s) - \bar{Z}_W(\beta, s)\} Y_{ki}(s) e^{\beta_0^T A_{ki}(s)} \lambda_0(s) ds, \end{aligned}$$

in which the second term becomes 0 when evaluated at $\beta = \beta_0$. Therefore, we have

$$\begin{aligned} K^{-1/2}U(\beta_0, \tau, W) = &K^{-1/2} \sum_{k=1}^{K} \sum_{i=1}^{n_0} \int_0^\tau \xi_{ki} W_{ki}(s) A_{ki}(s) dM_{ki}(s) \\ &- K^{-1/2} \int_0^\tau \bar{Z}_W(\beta_0, s) d\bar{M}(s), \end{aligned}$$

where $\bar{M}(s) = \sum_{k=1}^{K} \sum_{i=1}^{n_0} \xi_{ki} W_{ki}(s) M_{ki}(s)$. Due to the possible dependence among observations within the same cluster, we cannot directly use martingale theory to prove the asymptotic normality of $K^{-1/2}U(\beta_0, \tau, W)$. Alternatively, we first show that the quantity $K^{-1/2} \int_0^\tau \bar{Z}_W(\beta_0, s) d\bar{M}(s)$ is asymptotically equivalent to $K^{-1/2} \int_0^\tau \bar{z}(\beta_0, s) d\bar{M}(s)$.

We can show that

$$\|\bar{Z}_W(\beta_0, t) - \bar{z}(\beta_0, t)\| = o_p(1). \tag{4.10}$$

To see this, we have

$$\left\| \frac{S_W^{(1)}(\beta_0, s)}{S_W^{(0)}(\beta_0, s)} - \frac{s_W^{(1)}(\beta_0, s)}{s_W^{(0)}(\beta_0, s)} \right\| = \left\| \frac{S_W^{(1)}(\beta_0, s) s_W^{(0)}(\beta_0, s) - s_W^{(1)}(\beta_0, s) S_W^{(0)}(\beta_0, s)}{S_W^{(0)}(\beta_0, s) s_W^{(0)}(\beta_0, s)} \right\|.$$

Assumption 4.3.4 guarantees that the denominator is asymptotically bounded away from 0. The numerator

$$\begin{aligned}&\|S_W^{(1)}(\beta_0,s)s_W^{(0)}(\beta_0,s)-s_W^{(1)}(\beta_0,s)S_W^{(0)}(\beta_0,s)\|\\ \leq&\|s_W^{(0)}(\beta_0,s)\|\cdot\|S_W^{(1)}(\beta_0,s)-s_W^{(1)}(\beta_0,s)\|+\|s_W^{(1)}(\beta_0,s)\|\cdot\|s_W^{(0)}(\beta_0,s)-S_W^{(0)}(\beta_0,s)\|.\end{aligned}$$

The right hand side (RHS) converges in probability to 0, by assumption 4.3.2. By lemma A.1 in Spiekerman and Lin (1998), we have

$$\left\|K^{-1/2}\int_0^\tau\{\bar{Z}_W(\beta_0,s)-\bar{z}(\beta_0,s)\}d\bar{M}(s)\right\|\to_p 0,$$

which implies that

$$K^{-1/2}\int_0^\tau\{\bar{Z}_W(\beta_0,s)-\bar{z}(\beta_0,s)\}d\bar{M}(s)\to_p 0.$$

The desired asymptotic equivalence is shown. Now we have $K^{-1/2}U(\beta_0,\tau,W)$ is asymptotically equivalent to

$$K^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\xi_{ki}\int_0^\tau W_{ki}(s)\{A_{ki}(s)-\bar{z}(\beta_0,s)\}dM_{ki}(s).$$

By multivariate central limit theorem, as $K\to\infty$, $K^{-1/2}U(\beta_0,\tau,W)$ converges to a zero mean Gaussian process with covariance matrix $G_U(\beta_0)$, where

$$G_U(\beta_0)=var\left\{\sum_{i=1}^{n}\xi_{1i}\int_0^\tau W_{1i}(s)\{A_{1i}(s)-\bar{z}(\beta_0,s)\}dM_{1i}(s)\right\}$$

□

We are now in the position to prove Theorem 4.3.2.

Proof. We decompose $K^{-1}\mathcal{I}(\beta_0, \tau)$ by

$$K^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_0}\xi_{ki}\int_0^\tau W_{ki}(t)\left[\frac{S_W^{(2)}(\beta_0,t)S_W^{(0)}(\beta_0,t) - \{S_W^{(1)}(\beta_0,t)\}^{\otimes 2}}{S_W^{(0)}(\beta_0,t)^2}\right]dM_{ki}(t)$$
$$+\int_0^\tau\left[\frac{S_W^{(2)}(\beta_0,t)S_W^{(0)}(\beta_0,t) - \{S_W^{(1)}(\beta_0,t)\}^{\otimes 2}}{S_W^{(0)}(\beta_0,t)^2}\right]S_W^{(0)}(\beta_0,t)\lambda_0(t)dt$$

The elements in the first term are local square integrable martingale which can be shown to converge to zero in probability. We now prove the second term $\int_0^\tau V_W(\beta_0,t)S_W^{(0)}(\beta_0,t)dt$ converge to the fixed matrix $G_W(\beta_0)$. Specifically,

$$\begin{aligned}
&\left\|\int_0^\tau V_W(\beta_0,t)S_W^{(0)}(\beta_0,t)dt - G_W(\beta_0)\right\| \\
=&\left\|\int_0^\tau \{V_W(\beta_0,t)S_W^{(0)}(\beta_0,t) - v(\beta_0,t)s_W^{(0)}(\beta_0,t)\}dt\right\| \\
\leq&\left\|\int_0^\tau \{V_W(\beta_0,t)S_W^{(0)}(\beta_0,t) - v(\beta_0,t)S_W^{(0)}(\beta_0,t)\}dt\right\| \\
+&\left\|\int_0^\tau \{v(\beta_0,t)S_W^{(0)}(\beta_0,t) - v(\beta_0,t)s_W^{(0)}(\beta_0,t)\}dt\right\|.
\end{aligned}$$

Under the regularity assumptions, the two terms in the last inequality converges to 0 in probability, respectively. Combining the Taylor expansion (4.9), theorem 4.3.1, lemma 4.3.1, we conclude the proof by applying Slutsky's theorem,

$$K^{1/2}(\hat{\beta}_W - \beta_0) \to_d N(0, G_W(\beta_0)^{-1}G_U(\beta_0)G_W(\beta_0)^{-1})$$

□

The variance-covariance matrix of $\hat{\beta}_W$ can be estimated by

$$\hat{G}_W(\hat{\beta}_W)^{-1}\hat{G}_U(\hat{\beta}_W)\hat{G}_W(\hat{\beta}_W)^{-1},$$

where $\hat{G}_U(\cdot)$ and $\hat{G}_W(\cdot)$ are obtained by replacing unknown quantities in $G_U(\cdot)$ and $G_W(\cdot)$ by their finite sample estimates, respectively.

4.4. Simulation

We conducted extensive simulation studies to investigate the ability of marginal structural Cox model in yielding average treatment effect estimates that are consistent with those in randomized trials. For better illustration, we used the scenario where patients are clustered within doctors.

4.4.1 Covariates and Correlated Failure Time

We generated three covariates that are considered 'important': (1) patient's age $X_1 \sim Uniform(30, 70)$; (2) indicator of severe pre-existing conditions X_2 following a Bernoulli distribution with $p = 0.3$; (3) indicator of possessing a comprehensive health insurance X_3 which follows a Bernoulli distribution with $p = 0.8$ if aged above 55, or a Bernoulli distribution with $p = 0.2$ if otherwise. They were deemed important in the sense that they were related to both treatment history and outcome, thus, were confounders. Besides the three confounders, we generated 10 mutually independent nuisance covariates, denoted by $V_1, \ldots, V_{10}$, each following a standard normal distribution. Though not related to either treatment history or outcome, they were included in constructing the weights for the marginal structural Cox model.

We designed two mechanisms from which experimental/control treatment Z_{ki} (1 for

experimental, 0 for control) is assigned. The first assignment mechanism was based on a logistic model with cluster-level random effect. Specifically, it follows a logistic model

$$logit(P(Z_{ki} = 1)) = -1 + 0.05X_{1,ki} - X_{2,ki} + 0.8X_{3,ki} + \eta_k.$$

The cluster-level random effect η_k follows a standard normal distribution and measures the doctor's overall tendency to assign treatment.

In practice, such mechanism may often be unrealistic because few doctor would assign treatment based on a logistic model. Instead, a doctor uses his own discretion and professional experience to determine whether a patient is too old or too sick to receive a potentially aggressive experimental treatment. Such decision procedures can be extremely complicated and definitely vary from doctor to doctor. Therefore, we need another mechanism that can better emulate the procedure. In our second mechanism, we introduced two more doctor-specific random effects in addition to η_k: the random effect ξ_k that reflected the doctor's judgment on the suitability of assigning treatment based on the patient's age, and μ_k which represented random variation of chances in assigning treatment. We assume that $\xi_k \sim Unif(-3, 3)$ and $\mu_k \sim Unif(-0.1, 0.1)$. Treatment is assigned by the following rules: if $X_{3,ki} = 1$ and $\eta_k \geq 0.4$, $Z_{ki} = Bernoulli(0.8 + \mu_k)$. Otherwise,

$$Z_{ki} = \begin{cases} Bernoulli(0.7 + \mu_k) & \text{, if } X_{1,ki} \leq 55 + \xi_k \text{ and } X_{2,ki} = 1 \\ Bernoulli(0.8 + \mu_k) & \text{, if } X_{1,ki} > 55 + \xi_k \text{ and } X_{2,ki} = 1 \\ Bernoulli(0.3 + \mu_k) & \text{, if } X_{1,ki} \leq 55 + \xi_k \text{ and } X_{2,ki} = 0 \\ Bernoulli(0.2 + \mu_k) & \text{, if } X_{1,ki} > 55 + \xi_k \text{ and } X_{2,ki} = 0 \end{cases}$$

We hereafter refer to the treatment assignment scheme via logistic model as scheme 1, and the complicated scenario as scheme 2.

We assumed that the marginal distribution of failure time T_{ki} has failure rate

$$\lambda_0(t)exp\{\beta_0^T W_{ki}\},$$

where β_0 was the true regression parameter vector and W_{ki} was the set of covariates. Correlated failure time data within doctor k were generated from the Clayton-Cuzick model (Clayton and Cuzick 1985), in which the joint survival function of $T_k = (T_{k1}, \ldots, T_{kn_k})^T$ had the form:

$$S(t_{k1}, \ldots, t_{kn_k}|W_{k1}, \ldots, W_{kn_k}) = \left\{ \sum_{i=1}^{n_k} exp(\frac{\int_0^{t_{ki}} \lambda_0(t) e^{\beta_0^T W_{ki}} dt}{\theta}) - (n_k - 1) \right\}^{-\theta}. \tag{4.11}$$

The positive parameter θ measured the strength of correlation among $(T_{k1}, \ldots, T_{kn_k})^T$. θ was related to Kendall's τ_θ in the way that $\tau_\theta = 1/(2\theta+1)$ when there is no censoring. The smaller θ was, the larger the Kendall's τ_θ, hence the stronger the correlation. We performed two types of simulation studies using two different failure time generating mechanisms. They are described in the following two sections.

4.4.2 Binary Time-independent Treatment

We first investigate the performance of marginal structural Cox model in the simplest situation where there is only one binary treatment Z. We assumed that, for subject i of doctor k, the failure time T arises from the marginal hazard model

$$\lambda_{ki}(t) = 0.5exp\{\beta Z_{ki} + 0.025X_{1,ki} + 0.25X_{2,ki} - 0.25X_{3,ki}\}. \tag{4.12}$$

Correlated failure times were simulated via Clayton-Cuzick model (4.11). By setting Z_{ki} to either 0 or 1, we obtained pairs of counter-factual failure times $(T_{0,ki}, T_{1,ki})$ to rep-

resent the subjects' potential outcomes had he/she been assigned to a specific treatment at baseline, in the absence of censoring. To emulate a randomized trial, we assigned the subjects to the experimental group by $Bernoulli(0.5)$. Depending on the assigned treatment group, corresponding value of the pair $(T_{0,ki}, T_{1,ki})$ was extracted. Right censoring time was then simulated from $Uniform(0, r)$ distribution. Data from an observational study was generated in a similar fashion, except that the treatment group was assigned according to the two schemes described in section 4.4.1. During this process, covariates X_1, X_2, X_3 were directly related to both potential outcome T and assigned treatment Z, therefore, were confounders.

On the generated observational data, we estimate the marginal treatment effect by fitting a model analogical to intent-to-treat Cox model.

$$\lambda_{ki}(t) = \lambda_0(t) exp\{\gamma Z_{ki}\} \tag{4.13}$$

Model (4.13) compares the relative risk between experimental and control groups, regardless of whether the subjects subsequently stop or initiate another therapy. In other words, it compares treatment initiators and non-initiators. Therefore, model (4.13) can be viewed as the observational equivalence of its intent-to-treat analysis counterpart. Estimation is done via inverse probability weighting. Weights at baseline were estimated by a mixed effect logistic model whose covariates were X_1, X_2, X_3 plus the nuisance covariates $V_1, \ldots, V_{10}$. Since model (4.13) is different from the data-generating model (4.12), the average treatment hazard ratio γ would differ from β, the conditional hazard ratio. As was stated in Gail et al. (1984) and Austin et al. (2007), in general we have $|\gamma| \leq |\beta|$ if $\beta \neq 0$. Ideally, we would integrate X_1, X_2, X_3 out of model (4.12) to obtain the true value. This may be infeasible if the conditional probability $f(Z|X_1, X_2, X_3)$ is complicated, e.g. scheme 2 in section 4.4.1. Alternatively, we elected to find the true value via simulation. Specifically, we generated a large data set (5000 doctors, each with 20 patients), randomly assigned

treatment from $Bernoulli(0.5)$ and obtained an estimate of average treatment hazard ratio. We repeated this procedure 500 times and referred to average hazard ratio as the true value.

We carried out the simulation by setting β in (4.12) to -2 or 0. The corresponding true values were -1.92 and 0.01, respectively. The number of doctors we considered were 100, 200 and 300. Each doctor had 20 patients assigned to him. Parameter θ in (4.11) was set to 1 so the patients assigned to the same doctor were moderately correlated. Values of right censoring parameter r were selected to produce about 85% censoring.

Results based on 500 simulations are presented in table 4.1. We computed the bias, empirical standard deviation (ESD), average of standard error (ESE) and coverage rate (CR) of 95% confidence intervals. We fitted the unadjusted LWA model (Lee et al. 1992) that does not address confounding by indication. The results showed that this method performed reasonably well under treatment assignment scheme 1, with unbiased estimates and good 95% CI coverage rates. However, it yielded biased estimates under the complicated scheme 2, and the coverage rates became lower as the number of clusters increased. On the other hand, estimators from MSCM (4.13) was approximately unbiased, even under the complicated treatment assignment mechanism 2. The ESEs were very close to the ESDs, indicating good approximation of the covariance matrix estimator. The coverage rates of 95% confidence intervals were satisfactory. The emulated RCT estimators was also approximately unbiased and were close to their MSCM counterparts. Under certain setups, marginal structural Cox model had a slightly higher variability, which was likely due to inverse probability weighting.

4.4.3 Primary Treatment, with a Possibility of Secondary Treatment

In both randomized trials and observational studies, failure to adhere to the treatment of primary interest often introduces great complexity to statistical analysis. In many cases, intent-to-treat analysis (or per-protocol analysis, for observational studies) are carried out, completely ignoring the adherence to the treatment that was initiated. On the other hand, an as-treated analysis classifies the subjects according to the actual treatment they received, as opposed to the treatment they initiated at baseline. As-treated analysis can take flexible forms. For example, the model can incorporate a time-dependent treatment group indicator $A(t)$. This is often referred to as 'current versus never users' comparison. In this section, we considered another important variation of as-treated analysis that accommodates secondary treatment.

Secondary treatment is prevalent in both randomized trials and observational studies. After a subject initiates a primary treatment at randomization/baseline, he or she later may have to start a secondary treatment due to deteriorating condition or other complications. An intent-to-treat analysis will fail to tease out the effect of primary treatment, which is usually of major interest, since it is confounded by secondary treatment effect. Marginal structural Cox models has been utilized to address this issue under randomized trial setting (Yamaguchi and Ohashi 2004, Zhang and Wang 2012). The model is defined using counter-factual framework. Let T_z^p denote the failure time of a subject if, possibly contrary to the fact, the subject received treatment z at baseline and initiated the secondary treatment at time p. We let $p = \infty$ if secondary treatment is never initiated. For each subject, this is an infinite-dimensional counter-factual since secondary treatment can be initiated at an arbitrary time point. We assumed a marginal structural Cox model

$$\lambda_{T_z^p}(t) = \lambda_0(t) exp\{\theta_1 z + \theta_2 q(t)\}, \tag{4.14}$$

where $q(t)$ is the time-dependent indicator of secondary treatment with $q(t) = I(t \geq p)$ and we have $a(t) = \{z, q(t)\}^T$. $\lambda_{T_z^p}(t)$ is the hazard function for T_z^p and $\lambda_0(t)$ is the unspecified baseline hazard function. In this simulation, we assumed that the failure times followed an exponential distribution. Therefore, we have a constant baseline hazard λ_0,

$$\lambda_{T_z^p}(t) = \lambda_0 exp\{\theta_1 z + \theta_2 q(t)\}.$$

The survival function of T_z^p is $exp\{-\int_0^t \lambda_0 exp\{\theta_1 z + \theta_2 q(s)\}ds\}$. Let $u \sim Uniform(0,1)$, by probability integral theorem, we generate T_z^p by solving t from

$$exp\{-\int_0^t \lambda_0 exp\{\theta_1 z + \theta_2 q(s)\}ds\} = u.$$

For $t \leq p$, we have $a(s) = 0$ for $s \leq t$ and obtain $T_z^p = -\frac{log(u)}{\lambda_0 exp\{\theta_1 z\}}$. For $t > p$, the equation became

$$\int_0^p exp\{\theta_1 z\}ds + \int_p^t exp\{\theta_1 z + \theta_2\}ds = -\frac{log(u)}{\lambda_0}.$$

We then have

$$\begin{aligned} T_z^p &= p + [-\frac{log(u)}{\lambda_0 exp\{\theta_1 z\}} - p]/exp(\theta_2) \\ &= p + [T_z^\infty - p]/exp(\theta_2) \end{aligned}$$

The derivation suggested that we could first generate the failure time without secondary treatment. Then if the subject initiated secondary treatment at moment p, his residual survival after p was prolonged by the factor $exp(\theta_2)$. By following these two steps, the true

values are identical to θ_1 and θ_2 if we fit the MSCM (4.14). Therefore, we do not need to obtain true values by emulating randomized trials repeatedly from a huge counter-factual data.

We now describe the details about generating the counter-factual dataset. First, we generated the counter-factual pair (T_0^∞, T_1^∞), the potential failure times if secondary treatment were never initiated. We assumed the failure times followed an exponential distribution with marginal hazard function

$$\lambda_{T_Z^\infty}(t) = \lambda_0 exp\{\theta_1 Z\}.$$

Correlated failure times within the same doctor were simulated from Clayton-Cuzick model (4.11). When generating (T_0^∞, T_1^∞) using probability integral theorem, the standard uniform random seed u required some special manipulation. We introduced two independent standard normal random variables X_4 and ϵ and a constant a between 0 and 1. Define $y = a \cdot X_4 + \sqrt{1-a^2} \cdot \epsilon$. Obviously, y was still standard normal due to the independence between X_4 and ϵ. Hence, $u = \Phi(y)$, where $\Phi(\cdot)$ was the standard normal cumulative distribution function, followed standard normal distribution. Indirectly, X_4 was related to the failure times (T_0^∞, T_1^∞) and the strength of dependence was determined by the value of a: the larger the a was, the stronger the dependence. Similarly, we could relate more covariates to the failure times.

The second stage was to generate the time P when secondary treatment was initiated from the model $\lambda_P(t) = \lambda_{0P}(t) exp\{\gamma_1 Z + \gamma_2 X_4\}$. For $Z = 0, 1$, if $P \geq T_Z^\infty$, then the secondary treatment was never initiated and $T_Z^P = T_Z^\infty$. Otherwise, we have $T_Z^P = P + (T_Z^\infty - P) exp\{-\theta_2\}$. For each subject, we have the counter-factual pair $(T_{0,ki}^P, T_{1,ki}^P)$. We emulated the case of randomized trials and observational studies in the same way as in section 4.4.2.

The covariate X_4 introduced in this section was indirectly related to the failure time. It was also related to treatment history $A(t)$ since the time to initiate secondary treatment P was directly dependent upon X_4. Therefore, X_4 was a confounder.

Most simulation setups were similar to those in section 4.4.2. In addition, $\lambda_{0P}(t)$ was selected so that about 40% of subjects initiated secondary treatment. Results based on 500 simulations are presented in table 4.2. As expected, emulated randomized trials gave both unbiased point estimates and good confidence interval coverage rates. Results also demonstrated that marginal structural Cox model was able to yield unbiased estimates for both primary and secondary treatments. In general, the ESEs were fairly close to the ESDs for both θ_1 and θ_2. We also noticed that there were a few outliers among θ_2 estimates, possibly due to the extreme values in the estimated weights. The coverage rates of 95% confidence intervals were satisfactory.

4.5. Data Analysis

To illustrate our method, we implemented the proposed method to a data set from the INSPIRIS study. Starting from January 2010, INSPIRIS Inc. started to offer a home visiting health care program in selected communities in Michigan. The aim of this home visiting health care program was to identify symptoms at an earlier stage and to provide proper medical precautions. The investigators were interested in whether the program can improve the quality of life and reduce health insurance claim payments. The enrollment of the program was not offered randomly. Instead, at the beginning of each month, investigators examined the medical records during the past twelve-month period and decided the program eligibility. If a subject had incurred a large amount of claim payment, s/he was more likely to be offered the program. Other factors relevant to program eligibility included number of hospitalization and emergency room visits, and disease history. Geographical area also played a role in determining the eligibility. Therefore, subjects living

in vicinity form clusters and are potentially correlated. The program was offered to 1,082 participants and claim data are available on 10,712 non-participants.

The response in our analysis was the time to the first emergency room (ER) visit after January 1, 2010, denoted by T. Since the program eligibility was evaluated monthly, we postulate a marginal structural Cox model for T:

$$\lambda_T(t) = \lambda_0(t)exp\{\beta \cdot a(t)\}, \tag{4.15}$$

where $a(t)$ is the time-dependent program membership indicator. Therefore, β is interpreted as the log hazard ratio between current INSPIRIS participants and non-participants. Subjects were followed until they had an emergency room visit, loss to follow-up, or administrative censoring date July 1, 2011, whichever came first. We set the administrative censoring date because claim data became fairly sparse after that date. We used the counting process style input and broke the follow-up period into month-long intervals. We also extracted disease and other medical history information from the claim data, using the clinically modified International Classification of Diseases (ICD-9-CM) codes. A full list of the indicators is in Table 4.3. For the cluster information, we used five digit zip code. The 10,183 subjects in our data formed 374 clusters with sizes ranging from 2 to 309.

In order to estimate the weights used in inverse probability weighting, we fitted a logistic regression model with random intercept on data in each month-long interval. The model adjusted for the number of hospitalization/ER visit, total claim payment and disease indicators during the past twelve-month period, as well as several baseline factors including gender and age on January 1, 2010. The hazard ratio estimate in MSCM (4.15) was 0.379 (95% CI: 0.336 to 0.427), which indicated that the INSPIRIS program is helpful in reducing ER visit. There are some limitations for this analysis. As we mentioned before, one critical assumption for MSCM is that there is no unmeasured confounding (assump-

tion 4.2.3). With these claim data, it is very likely that there could be some important confounding factors which are not measured.

4.6. Discussion

Large scale electronic medical record databases have seen rapidly growing popularity in recent years. One largest challenge in estimating average treatment effect using EMR data is to eliminate the confounding by indication. Meanwhile, it is common to have clustered subjects whose responses are correlated in EMR data. In this paper, we proposed a marginal structural Cox model approach that can handle time-to-event data with cluster-level random effect and correctly estimate the average treatment effect. We formulated the model using counter-factual arguments. Parameters can be estimated using inverse probability weighting technique. We proved its asymptotic properties via martingale theory. We implemented our method in both simulations and a real large scale observational claim data.

Claim data generally do not have mortality information, unless linked to external sources such as hospital administrative data and clinical data (Pine et al. 1997). There were attempts made to predict mortality using the ICD-9-CM codes recorded in claim data (Iezzoni et al. 1995). In the absence of mortality data, investigators sometimes will use alternative endpoints. Some endpoints such as hospitalization and ER visit are recurrent. Our marginal structural Cox model approach was based on the AG (Andersen and Gill 1982) model for terminal events such as death. While the AG method could potentially handle recurrent events, a more suitable approach that we could develop in the future is based on the rate models by Pepe and Cai (1993) for recurrent events.

When analyzing claim data, investigators are usually interested in the claim payment. In many cases, including a parallel analysis of the INSPIRIS data, the claim payment is

considered as a continuous endpoint and linear models are usually employed. A more proper approach to analyze claim payments is to utilize methods concerning medical costs. Because a patient who accumulates costs over time at relatively higher rates tends to generate larger cumulative costs at both the survival time and censoring time, the cumulative cost at both times are positively correlated. Standard survival analysis methods may be invalid because of the informative censoring. To this end, Lin (2000a) proposed a proportional means regression model to handle such data. In claim data, geological information may also introduce cluster level heterogeneity. This is because medical cost for subjects living in vicinity may have similar socio-economic status and access to health care. In contrast, such characteristics can vary significantly across clusters. To our best knowledge, there is no statistical procedure for analyzing medical cost data that can address cluster level heterogeneity and confounding by indication simultaneously. A possible future work is to fill this methodological gap.

Table 4.1: *Simulation Results: Binary Time-independent Treatment*

			Unadjusted				RCT				MSCM			
Beta	Scheme	#Doctors	Bias	ESD	ESE	CR	Bias	ESD	ESE	CR	Bias	ESD	ESE	CR
-2	1	100	-0.019	0.145	0.143	0.940	-0.014	0.141	0.138	0.950	-0.023	0.146	0.143	0.940
		200	-0.008	0.101	0.102	0.940	-0.003	0.104	0.098	0.940	-0.012	0.102	0.102	0.940
		300	-0.011	0.078	0.083	0.968	0.000	0.078	0.080	0.962	-0.015	0.079	0.084	0.960
	2	100	0.097	0.138	0.140	0.896	-0.004	0.138	0.138	0.944	0.038	0.148	0.147	0.944
		200	0.086	0.100	0.100	0.858	-0.008	0.104	0.098	0.934	0.029	0.106	0.105	0.942
		300	0.096	0.085	0.082	0.766	0.000	0.084	0.080	0.946	0.041	0.089	0.086	0.944
0	1	100	-0.012	0.126	0.124	0.946	-0.012	0.112	0.115	0.956	-0.016	0.126	0.125	0.948
		200	-0.011	0.090	0.088	0.944	-0.013	0.085	0.082	0.940	-0.015	0.091	0.089	0.946
		300	-0.010	0.068	0.072	0.952	-0.009	0.065	0.067	0.958	-0.015	0.069	0.073	0.954
	2	100	0.090	0.114	0.116	0.882	-0.008	0.114	0.115	0.944	0.037	0.119	0.120	0.942
		200	0.074	0.085	0.082	0.848	-0.016	0.084	0.081	0.930	0.019	0.087	0.085	0.930
		300	0.082	0.068	0.067	0.764	-0.010	0.069	0.067	0.960	0.029	0.071	0.069	0.930

NOTE: ESD, empirical standard deviation; ESE, average standard error estimator; CR, estimated standard error coverage rate of the nominal 95% confidence intervals

Table 4.2: *Simulation Results: With Possible Secondary Treatment*

				RCT				MSCM			
$(\theta_1, \theta_2)^T$	Scheme	#Doctor	Parameter	Bias	ESD	ESE	CR	Bias	ESD	ESE	CR
(-0.69, -0.2)	1	200	θ_1	-0.032	0.135	0.135	0.946	-0.017	0.168	0.157	0.932
			θ_2	0.004	0.162	0.148	0.950	-0.002	0.160	0.131	0.930
		250	θ_1	-0.027	0.112	0.119	0.954	-0.015	0.143	0.143	0.950
			θ_2	-0.008	0.149	0.134	0.944	0.007	0.132	0.121	0.940
		300	θ_1	-0.036	0.109	0.109	0.938	-0.015	0.138	0.138	0.950
			θ_2	0.029	0.130	0.129	0.926	0.009	0.137	0.117	0.936
	2	200	θ_1	-0.017	0.133	0.133	0.950	0.010	0.138	0.141	0.952
			θ_2	0.018	0.175	0.157	0.926	0.006	0.135	0.127	0.940
		250	θ_1	-0.020	0.111	0.116	0.952	-0.004	0.120	0.122	0.952
			θ_2	0.023	0.147	0.139	0.928	0.005	0.118	0.115	0.950
		300	θ_1	-0.030	0.104	0.111	0.942	0.001	0.114	0.116	0.950
			θ_2	0.024	0.134	0.122	0.932	0.002	0.119	0.109	0.936
(0, -0.2)	1	200	θ_1	0.032	0.129	0.126	0.942	0.000	0.125	0.131	0.956
			θ_2	0.009	0.129	0.120	0.938	-0.002	0.139	0.121	0.920
		250	θ_1	0.032	0.108	0.110	0.934	-0.011	0.112	0.118	0.952
			θ_2	-0.007	0.121	0.112	0.942	-0.014	0.115	0.110	0.946
		300	θ_1	0.034	0.107	0.102	0.930	0.003	0.112	0.114	0.956
			θ_2	-0.019	0.108	0.107	0.930	0.001	0.138	0.111	0.924
	2	200	θ_1	0.024	0.124	0.124	0.942	-0.006	0.122	0.125	0.948
			θ_2	-0.020	0.145	0.131	0.936	-0.005	0.126	0.119	0.942
		250	θ_1	0.023	0.107	0.107	0.942	0.002	0.102	0.108	0.952
			θ_2	-0.018	0.116	0.116	0.930	-0.007	0.109	0.106	0.942
		300	θ_1	0.030	0.099	0.103	0.958	-0.006	0.099	0.104	0.962
			θ_2	-0.017	0.104	0.103	0.932	-0.008	0.110	0.101	0.934

NOTE: ESD, sample standard deviation; ESE, average standard error estimator; CR, estimated standard error coverage rate of the nominal 95% confidence intervals

Table 4.3: *List of Disease and Medical History Indicators*

Alcohol status	Arthritis	Behavioral disorder	Cancer
Chronic obstructive lung disease	Chronic renal disease	Congestive heart failure	Coronary artery disease
Delirium	Dementia	Diabetes	Encephalopathy
Falls	Hip Fracture	Hypertension	Smoking status
Stroke	Substance abuse		

CHAPTER 5: MIXED EFFECT MODEL FOR CLUSTER-BASED PDS

5.1. Introduction

Observational studies play a key role in investigating the relationship between outcome and exposure and other covariates. All studies are conducted with a limited budget and the maximum study sizes are often restricted by the cost of the exposure ascertainment. In practice, there are many scenarios where all outcome data are available, and a sub-study is planned which requires retrospectively collecting additional key exposure information on a limited number of subjects. As a result, cost-effective study designs, especially biased sampling designs, have been investigated closely. Among the biased sampling designs, the most widely used one is the case-control design due to its efficiency and cost-effective feature (Anderson 1972, Prentice and Pyke 1979). The fundamental idea of case-control design is to over-represent the cases that are considered to be more informative in relating response and exposure.

Case-control design is suitable when the outcome of interest is binary. On the other hand, there are numerous situations where the outcome of interest is measured continuously to which case-control design does not naturally extended. In practice, one *ad hoc* solution is to dichotomize the outcome based on a pre-specified threshold. However, it is obvious that there will be a loss of information in the continuous outcome. Meanwhile, the results may be sensitive to the potentially subjective choice of cutoff. For continuous time-to-event data, Prentice (1986) proposed case-cohort studies. Outcome-dependent sampling (ODS) design (Zhou et al. 2002, Weaver and Zhou 2005) is a two-stage biased sampling design proposed for general continuous responses. It is ideal for studies that values of outcome Y are known for all subjects, but the exposure variable X may be expensive or diffi-

cult to ascertain. Assume that the domain of Y can be partitioned into K mutually exclusive and exhaustive strata by known constants $-\infty = a_0 < a_1 < \cdots < a_{K-1} < a_K = \infty$ and let the kth stratum be represented by $C_k = (a_{k-1}, a_k], k = 1, \ldots, K$. The data structure of the ODS sample consists of a first-phase simple random sample (SRS) of size n_0 and a second-phase simple random sample of size $n_1, \ldots, n_K$ from each of the K strata. The latter is referred to as 'supplementary sample'. Through sampling the response Y at its two distributional tails, the X-values to be observed are more likely to occur at its distributional tails as well if the true underlying distribution is linear. Linear model theory shows that the variance of parameter estimate is inversely proportional to the summed squares of observed X-values ($X^T X$). Therefore, having a sample that X-values are over-represented in its two distributional tails will be more informative than having a simple random sample in which X-values are evenly concentrated around its mean. For variations of ODS, see Zhou et al. (2011a), Qin and Zhou (2011) and Zhou et al. (2011c). In many studies, there may exist a continuous auxiliary variable W for X, which is available for all subjects in the full cohort. Intuitively, incorporating auxiliary information in W in biased sampling may lead to improved efficiency. To this end, outcome-auxiliary-dependent sampling (OADS) designs are proposed in Wang and Zhou (2010) and Zhou et al. (2011b). Suppose that W can be partitioned into J mutually exclusive and exhaustive strata by known constants $-\infty = b_0 < b_1 < \cdots < b_{J-1} < b_J = \infty$ and let the jth stratum be represented by $B_j = (b_{j-1}, b_j], j = 1, \ldots, J$. Then the population can be partitioned into $T = K \times J$ strata on the domain of $Y \times W$. The outcome-auxiliary-dependent sample also consists of two components: an overall SRS and supplementary samples from each of the T strata.

OADS may become implausible when there are a large number of strata for the auxiliary variable or the combination of several auxiliary variables. The probability-dependent sampling (PDS) design, proposed in (Zhou et al. 2014), is well-suited in this scenario. The rationale of PDS is to stratify the response-auxiliary domain by a single 'score', that is, a probability. Another appealing feature of PDS is that it allows investigators to over-sample

the two tails of X distribution without the prior knowledge that the relationship between Y and X is linear. Suppose that the domain of the exposure X is partitioned into three mutually exclusive intervals: $(-\infty, x_L] \cup (x_L, x_U] \cup (x_U, \infty)$. Like ODS and OADS, an SRS is drawn from the population in the first stage. Before the second stage supplementary sampling, a model for $E(X|Y, Z)$ is fitted using the first-phase SRS. On the basis of this model, the chances of a new subject's X conditional on $Y = y$ and $Z = z$, will be in $(-\infty, x_L]$ and (x_U, ∞) are predicted by $\hat{\phi}_1(y, z) = \hat{pr}(X < x_L|Y, Z)$ and $\hat{\phi}_3(y, z) = \hat{pr}(X > x_U|Y, Z)$ respectively. Then supplementary samples are drawn from those whose X are more likely to fall on the distributional tails of X. For example, random samples can be drawn from those with $\hat{\phi}_1(y, z) > c_1$ or with $\hat{\phi}_3(y, z) > c_3$, where c_1 and c_3 are thresholds.

In medical studies, it is very common that subjects from the same clinic or community form clusters. The Collaborative Perinatal Project (CPP, Niswander and Gordon (1972)), for example, is an epidemiological study where investigators were interested in studying in utero exposure to polychlorinated biphenyls (PCBs) in relation to various health outcomes that include neurodevelopmental abnormalities, among children born in the CPP study. The PCB levels were measured retrospectively by analyzing the preserved third-trimester blood serum specimens, which was costly. CPP participants were enrolled through 12 university-affiliated medical clinics, with the centers contributing unequal numbers of subjects. Therefore, subjects within the same clinic may be correlated and it is important that the clinic-specific random effect can be properly addressed in statistical analysis. Despite of the advances in ODS-like designs with independent subjects, the literature on such designs with correlated subjects is very limited. Xu and Zhou (2012) considered a linear mixed effect model (Fitzmaurice et al. 2004) for cluster-based OADS design. Schildcrout et al. (2013) proposed an outcome vector dependent sampling design for longitudinal continuous response data, based on summary statistics of individual longitudinal trajectories. To our best knowledge, there is a gap in methodology that addresses cluster level heterogeneity in PDS designs.

In this paper, we propose a cluster-based probability-dependent sampling design. The rest of this paper is structured as follows. In Section 5.2.1, we describe the design and data structure. The estimation and inference procedures based on a profile likelihood function are presented in Section 5.2.2. We performed simulation studies to assess the performance of our estimator in Section 5.3. In Section 5.4, we illustrate our method by implementing it on the CPP data. We conclude with some discussion in Section 5.5.

5.2. Design and Semiparametric Inference

5.2.1 Design and Data Structure

Let Y denote the continuous outcome, (X, Z) denote the vector of covariates where X is the expensive scalar exposure and Z is the vector of covariates that are available for all subjects. We use i to index clusters and j to index the subjects within the cluster. We assume a linear mixed effect model for Y_{ij} given (X_{ij}, Z_{ij})

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \eta_i + \epsilon_{ij}, \tag{5.1}$$

where $\eta_i \sim N(0, \sigma_\eta^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. We assume that η_i and ϵ_{ij} are independent, analogous to the ordinary linear mixed effect model. The parameter vector $\beta = (\beta_0, \beta_1, \beta_2, \sigma_\eta^2, \sigma_\epsilon^2)^T$. Let $x_L < x_U$ be known constants that partition the domain of X into three mutually exclusive intervals $A_1 \cup A_2 \cup A_3 = (-\infty, x_L] \cup (x_L, x_U] \cup (x_U, \infty)$.

The sampling has two stages. In the first stage of PDS, we draw an SRS of size n_0 from m_0 clusters and their X-values are ascertained. On the first-stage SRS, a model of $E(X|Y, Z)$ is fitted. We may use linear models, logistic models or non-parametric kernel methods. For the subjects outside the first-stage SRS, we can obtain the predicted values of the following two conditional probabilities of X falling into the lower/upper stratum given

Y and Z: $\phi_1(Y,Z) = pr(X \in A_1|Y,Z)$ and $\phi_3(Y,Z) = pr(X \in A_3|Y,Z)$, denoted by $\hat{\phi}_1(Y,Z)$ and $\hat{\phi}_3(Y,Z)$. The second-phase supplementary sampling is conducted based on the predicted probabilities $\hat{\phi}_1(Y,Z), \hat{\phi}_3(Y,Z)$. An SRS of size $n_k, k = 1,3$ is drawn from each stratum that $\hat{\phi}_k(Y,Z)$ is greater or equal to pre-specified threshold c_k, e.g., 80%. The number of clusters in supplementary sample are m_1 and m_3 respectively. The total sample size $n = \sum_k n_k = \sum_k \sum_{i=1}^{m_k} n_{ki}$ and the total number of clusters $m = m_0 + m_1 + m_3$. The data structure for the proposed cluster-based PDS is as follows.

First-stage SRS,

$$\{(Y_{0ij}, X_{0ij}, Z_{0ij})\}, i = 1, \cdots, m_0, j = 1, \cdots, n_{0i};$$

and second-stage supplementary sample,

$$\{(Y_{1ij}, X_{1ij}, Z_{1ij}) : pr(X_{1ij} \in A_1|Y_{1ij}, Z_{1ij}) \geq c_1\}, i = 1, \cdots, m_1, j = 1, \cdots, n_{1i};$$
$$\{(Y_{3ij}, X_{3ij}, Z_{3ij}) : pr(X_{3ij} \in A_3|Y_{3ij}, Z_{3ij}) \geq c_3\}, i = 1, \cdots, m_3, j = 1, \cdots, n_{3i}.$$

5.2.2 Estimation and Asymptotic Results

Let $G(X,Z)$ and $g(X,Z)$ denote the joint CDF and PDF of (X,Z), respectively. Note that,

$$\begin{aligned} f(Y_{ij}, X_{ij}, Z_{ij}, \eta_i) &= f(Y_{ij}|X_{ij}, Z_{ij}, \eta_i) \cdot h(\eta_i|X_{ij}, Z_{ij}) \cdot g(X_{ij}, Z_{ij}) \\ &= f(Y_{ij}|X_{ij}, Z_{ij}, \eta_i) \cdot h(\eta_i) \cdot g(X_{ij}, Z_{ij}). \end{aligned}$$

The last equation is granted by assuming the independence between cluster-level and individual-level random effects. We also assume that the observations of (X,Z) within

the same cluster are independent. This is usually the case if the subjects are clustered within the same participating clinic, like in the CPP.

With known $\phi_k(Y_{kij}, Z_{kij})$, the likelihood function can be expressed as

$$L(\beta, G) = \left\{ \prod_{i=1}^{m_0} \int \prod_{j=1}^{n_{0i}} f_\beta(Y_{0ij}|X_{0ij}, Z_{0ij}, \eta) h(\eta) g(X_{0ij}, Z_{0ij}) d\eta \right\}$$
$$\times \left\{ \prod_{k=1,3} \prod_{i=1}^{m_k} \int \prod_{j=1}^{n_{ki}} f_\beta(Y_{kij}, X_{kij}, Z_{kij}, \eta | \phi_k(Y_{kij}, Z_{kij}) \geq c_k) d\eta \right\}, \qquad (5.2)$$

For $k = 1, 3$, using Bayes formula, we have

$$\begin{aligned} & f_\beta(Y_{kij}, X_{kij}, Z_{kij}, \eta_{ki} | \phi_k(Y_{kij}, Z_{kij} \geq c_k)) \\ =& \frac{f_\beta(Y_{kij}, X_{kij}, Z_{kij}, \eta_{ki}, I\{\phi_k(Y_{kij}, Z_{kij}) \geq c_k\})}{pr\{\phi_k(Y_{kij}, Z_{kij}) \geq c_k\}} = \frac{f_\beta(Y_{kij}, X_{kij}, Z_{kij}, \eta_{ki})}{pr\{\phi_k(Y_{kij}, Z_{kij}) \geq c_k\}} \\ =& f_\beta(Y_{kij}|X_{kij}, Z_{kij}, \eta_{ki}) h(\eta_{ki}) g(X_{kij}, Z_{kij}) \pi_k^{-1}, \end{aligned}$$

where

$$\pi_k = \int \int \int \int f(Y|X, Z, \eta) h(\eta) g(X, Z) I\{(Y, Z) : \phi_k(Y, Z) \geq c_k\} dY dX dZ d\eta.$$

Therefore, likelihood function $L(\beta, G)$ in (5.2) can be expressed as

$$\left\{ \prod_{i=1}^{m_0} \int \prod_{j=1}^{n_{0i}} f_\beta(Y_{0ij}|X_{0ij}, Z_{0ij}, \eta) h(\eta) g(X_{0ij}, Z_{0ij}) d\eta \right\}$$
$$\times \left\{ \prod_{k=1,3} \prod_{i=1}^{m_k} \int \prod_{j=1}^{n_{ki}} f_\beta(Y_{kij}|X_{kij}, Z_{kij}, \eta) h(\eta) g(X_{kij}, Z_{kij}) d\eta \right\} \prod_{k=1,3} \pi_k^{-n_k}, \qquad (5.3)$$

It is clear that maximizing (5.3) will inevitably involve addressing $G(X, Z)$, the joint

distribution of (X, Z). We propose a semiparametric likelihood method without specifying $G(X, Z)$. Note that for a fixed β, (5.3) is a biased sampling likelihood (Vardi 1982, 1985, Qin 1993). Let $p_{ij} = g(X_{ij}, Z_{ij})$. The log-likelihood is

$$\begin{aligned} l(\beta, \{p_{ij}\}, \pi_1, \pi_3) &= \left\{ \sum_{i=1}^{m} log\{ \int \prod_{j=1}^{n_i} f_\beta(Y_{ij}|X_{ij}, Z_{ij}, \eta) h(\eta) d\eta \} \right\} \\ &\quad + \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n_i} log(p_{ij}) - n_1 log(\pi_1) - n_3 log(\pi_3) \right\} \\ &= l_1(\beta) + l_2(\{p_{ij}\}, \pi_1, \pi_3), \end{aligned} \tag{5.4}$$

where $l_1(\beta)$ and $l_2(\{p_{ij}\}, \pi_1, \pi_3)$ denote the quantities in the first bracket and the second bracket, respectively.

The first step is to profile (5.4) over $\{p_{ij}\}$ by fixing (β, π_1, π_3) and to obtain the empirical likelihood function of $\{p_{ij}\}$ over all distributions whose support contains the observed values of X and Z. Since we still have independent components in the term $\sum_{i=1}^{m} \sum_{j=1}^{n_i} log(p_{ij})$, we can search for $\{\hat{p}_{ij}\}$ that maximize $l_2(\{p_{ij}\}, \pi_1, \pi_3)$ in (5.4), subject to following four constraints:

$$\begin{aligned} &\left\{ p_{ij} > 0; \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} = 1; \right. \\ &\sum_{i=1}^{m} \int \sum_{j=1}^{n_i} p_{ij} \left\{ \int f_\beta(Y|X_{ij}, Z_{ij}, \eta) h(\eta) I\{(Y, Z_{ij}) : \phi_1(Y, Z_{ij}) \geq c_1\} dY \right\} d\eta - \pi_1 = 0; \\ &\left. \sum_{i=1}^{m} \int \sum_{j=1}^{n_i} p_{ij} \left\{ \int f_\beta(Y|X_{ij}, Z_{ij}, \eta) h(\eta) I\{(Y, Z_{ij}) : \phi_3(Y, Z_{ij}) \geq c_3\} dY \right\} d\eta - \pi_3 = 0 \right\}. \end{aligned} \tag{5.5}$$

The four constraints rise from the idea in Owen (1988, 1990) that we need to consider only discrete distributions with jumps at each of the observed points.

We can use Lagrange multiplier argument to derive $\{\hat{p}_{ij}\}$ that maximizes (5.4) subject to the constraints (5.5). Specifically, we maximize the target function

$$
\begin{aligned}
& H(\beta, \{p_{ij}\}, \pi_1, \pi_3) \\
=& l_2(\{p_{ij}\}, \pi_1, \pi_3) + \rho(\sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} - 1) + n\sum_{k=1,3}\lambda_k\Big[\sum_{i=1}^{m}\times \\
& \int \sum_{j=1}^{n_i} p_{ij}\Big\{\int f_\beta(Y|X_{ij}, Z_{ij}, \eta)h(\eta)I\{(Y, Z_{ij}) : \phi_k(Y, Z_{ij}) \geq c_k\}dY\Big\}d\eta - \pi_k\Big],
\end{aligned}
\tag{5.6}
$$

where $\rho, \lambda_1, \lambda_3$ are Lagrange multipliers. Differentiate $H(\beta, \{p_{ij}\}, \pi_1, \pi_3)$ with respect to $\rho, \lambda_1, \lambda_3$ and $\{p_{ij}\}$, we get

$$
\begin{aligned}
\frac{\partial H}{\partial \rho} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} p_{ij} - 1 \\
\frac{\partial H}{\partial \lambda_k} &= n\sum_{i=1}^{m}\int\sum_{j=1}^{n_i} p_{ij}\Big\{\int f_\beta(Y|X_{ij}, Z_{ij}, \eta)h(\eta)I\{(Y, Z_{ij}) : \phi_k(Y, Z_{ij}) \geq c_k\}dY d\eta\Big\} - \pi_k \\
\frac{\partial H}{\partial p_{ij}} &= n\sum_{k=1,3}\lambda_k\Big\{\int\int f_\beta(Y|X_{ij}, Z_{ij}, \eta)h(\eta)I\{(Y, Z_{ij}) : \phi_k(Y, Z_{ij}) \geq c_k\}dY d\eta - \pi_k\Big\} \\
&\quad + \frac{1}{p_{ij}} + \rho
\end{aligned}
\tag{5.7}
$$

Set all derivatives to zero and solve the equation. Multiply each $\partial H/\partial p_{ij}$ by p_{ij} and sum over indices i, j, we can get $\hat{\rho} = n$. Plugging it back into $\partial H/\partial p_{ij}$, we obtain the maximizer $\hat{p}_{ij}$:

$$
n^{-1}\Big[1 + \sum_{k=1,3}\lambda_k\Big\{\int\int f_\beta(Y|X_{ij}, Z_{ij}, \eta)h(\eta)I\{(Y, Z_{ij}) : \phi_k(Y, Z_{ij}) \geq c_k\}dY d\eta - \pi_k\Big\}\Big]^{-1}. \tag{5.8}
$$

Replacing p_{ij} with (5.8) in (5.4), we have an estimated profile log-likelihood function

whose arguments are $(\beta, \pi_1, \pi_3, \lambda_1, \lambda_3)$. For unbiased sampling schemes, the true value of π_1 and π_3 are zero. But this is generally not the case with a biased sample. Therefore, to unify the notation, we center λ_1 and λ_3 by reparameterizing

$$\nu_1 = \lambda_1 - q_1/\pi_1,\ \nu_1 = \lambda_3 - q_3/\pi_3.$$

where $q_k = n_k/n$ for $k = 0, 1, 3$.

Define the following quantities: $\xi = (\beta^T, \pi_1, \pi_3, \nu_1, \nu_3)^T$,

$$F_k(X_{ij}, Z_{ij}) = \int\int f_\beta(Y|X_{ij}, Z_{ij}, \eta)h(\eta)I\{(Y, Z_{ij}) : \phi_k(Y, Z_{ij}) \geq c_k\}dY\,d\eta,$$

and

$$\Delta(X_{ij}, Z_{ij}) = q_0 + \frac{q_1}{\pi_1}F_1(X_{ij}, Z_{ij}) + \frac{q_3}{\pi_3}F_3(X_{ij}, Z_{ij}).$$

Replacing λ_k with ν_k in (5.8), the resulting reparameterized profile log-likelihood can be expressed as

$$\begin{aligned} l(\xi) &= l_1(\beta) - n_1 log(\pi_1) - n_3 log(\pi_3) \\ &\quad - \sum_{i=1}^{m}\sum_{j=1}^{n_i} log\left\{\Delta(X_{ij}, Z_{ij}) + \sum_{k=1,3}\nu_k\{F_k(X_{ij}, Z_{ij}) - \pi_k\}\right\} \end{aligned} \tag{5.9}$$

We replace $\phi_k(Y, Z_{ij})$ with $\hat{\phi}_k(Y, Z_{ij})$ in $F_k(X_{ij}, Z_{ij})$ and $\Delta(X_{ij}, Z_{ij})$ to obtain their estimated counterparts, denoted by $\hat{F}_k(X_{ij}, Z_{ij})$ and $\hat{\Delta}(X_{ij}, Z_{ij})$. The estimated version of

profile log-likelihood function has a form very similar to (5.9):

$$\hat{l}(\xi) = l_1(\beta) - n_1 log(\pi_1) - n_3 log(\pi_3) \\ - \sum_{i=1}^{m} \sum_{j=1}^{n_i} log\left\{ \hat{\Delta}(X_{ij}, Z_{ij}) + \sum_{k=1,3} \nu_k \{\hat{F}_k(X_{ij}, Z_{ij}) - \pi_k\} \right\} \tag{5.10}$$

We can obtain the maximum semiparametric empirical likelihood estimator $\hat{\xi}$ for (5.10) by using Newton-Raphson algorithm. We hereby present the asymptotic results of our cluster-based PDS estimator $\hat{\xi}$ and the detailed proof is given in the section 5.6.

Theorem 5.2.1. *Under the regularity conditions outlined in the section 5.6,* $\hat{\xi}$ *satisfying* $n^{-1}\hat{l}(\hat{\xi}) = 0$ *converges almost surely to the true value* $\xi_0 = (\beta_0, \pi_1, \pi_3, 0, 0)^T$. *In addition,*

$$n^{1/2}(\hat{\xi} - \xi_0) \xrightarrow{d} N(0, \Sigma(\xi_0)),$$

where $\Sigma(\xi_0) = [V(\xi_0)^{-1}]\Omega(\xi_0)[V(\xi_0)^{-1}]^T$.

Explicit forms of $V(\xi)$ and $\Omega(\xi)$ are given in the section 5.6. A consistent estimator of $\Sigma(\xi_0)$ is $[\hat{V}(\hat{\xi})^{-1}]\hat{\Omega}(\hat{\xi})[\hat{V}(\hat{\xi})^{-1}]^T$, where $\hat{V}(\xi)$ and $\hat{\Omega}(\xi)$ are obtained by replacing theoretical quantities with finite sample estimates in $V(\xi)$ and $\Omega(\xi)$.

5.3. Simulation

We evaluate the performance of the proposed estimator by extensive simulation studies. We assume that the continuous main exposure X follows a standard normal distribution. Its domain $\mathcal{X}$ can be partitioned into three mutually exclusive intervals: $\mathcal{X} = A_1 \cup A_2 \cup A_3$, where $A_1 = (-\infty, \mu_X - a*\sigma_X]$, $A_2 = (\mu_X - a*\sigma_X, \mu_X + a*\sigma_X]$ and $A_3 = (\mu_X + a*\sigma_X, \infty)$. We generate a covariate Z which equals X plus a standard normal random error. We

assume that Z is measured on all the subjects in the full study cohort. To generate clustered data, the response variable Y arises from a linear model with a random intercept:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \eta_i + \epsilon_{ij}, \tag{5.11}$$

in which footnote i indexes clusters and j indexes subjects within a cluster. The individual-level random effects ϵ_{ij} follow a standard normal distribution and are independent over i, j. Cluster level random effects η_i are normally distributed with mean 0 and variance σ_η^2. Therefore, the population intra-class correlation (ICC) equals $\sigma_\eta^2/(1+\sigma_\eta^2)$ and can be controlled by a single factor σ_η^2. Without loss of generality, we assume that the all clusters have the same size. For the full study cohort, we generated 50 clusters of size 80, resulting in 4000 subjects. In each independent simulation run, we compared the following three estimators.

The first estimator, denoted by $\hat{\beta}_F$, is based on a hypothetical scenario in which we can observe the X values on all subjects in the full study cohort. Although this can be achieved in the simulation study, we emphasize that this is implausible in practice. This estimator will serve as the benchmark. The second estimator, denoted by $\hat{\beta}_S$, is the linear mixed effect model estimator. It is based on simple random sample of size n from the full study cohort.

The third estimator, denoted by $\hat{\beta}_P$, is our proposed PDS estimator. To implement this estimator, we first draw a simple random sample of size $n_0 = 200$ from the full study cohort. We fit two mixed effect logistic regression models on this sample. One is for the event $X_{ij} \in A_1$, while the other is for the event $X_{ij} \in A_3$. Both models adjust for Y, Z and account for the cluster level random effect. We then obtain the predicted probabilities $\hat{pr}(X_{ij} \in A_k | Y_{ij}, Z_{ij}, \eta_i), k = 1, 3$ for those who outside the first phase simple random sample. Supplementary probability dependent sampling is then conducted at $a = 1, 1.5$.

We investigated two probability sampling thresholds $c = c_1 = c_3$: $\hat{\beta}_{P1}$ corresponds to $c = 85\%$, while $\hat{\beta}_{P2}$ corresponds to $c = 95\%$. The sample sizes in the two probabilistic tails are the same, that is, $n_1 = n_3 = 100$. The total sample size $n = 400$. True values of β_1 and β_2 are set to 2 and -0.5, respectively.

Results based on 1000 independent simulations are presented in Table 5.1. All three types of estimators are able to yield unbiased point estimates with satisfactory 95% confidence interval coverage rates. The means of estimated standard errors are close to their corresponding empirical standard errors, indicating good approximation of the variance estimator. For each estimator, we also computed its relative efficiency (RE) to the benchmark full cohort estimator $\hat{\beta}_F$. Specifically, RE is defined as the squared ratio of ESE for $\hat{\beta}_F$ to ESE for the estimator. $\hat{\beta}_F$ itself has RE equals 1 by definition. The larger the RE is, the more efficient the estimator is. In our simulations, the cluster-based PDS estimators are more efficient than $\hat{\beta}_S$ obtained from a simple random sample of same size. For example, when ICC equals 0.33, the relative efficiencies of estimator from a simple random sample are both 0.088 for β_1 and β_2. In contrast, the relative efficiencies of our PDS estimator with $a = 1$ and $c = 85\%$ are 0.137 and 0.153, respectively for β_1 and β_2. In general, we gain more efficiency when the ICC is smaller, that is, when cluster level random effect contributes less to the total variability. On the other hand, the performance of cluster-based PDS estimator appears to be insensitive to (1) how we classify the tails of X domain ($a = 1, 1.5$) and (2) probability sampling threshold ($c = 85\%, 95\%$).

5.4. CPP Data Analysis

We illustrate our method by using data from the Collaborative Perinatal Project. It is an epidemiological study where investigators were interested in studying in utero exposure to polychlorinated biphenyls (PCB) in relation to various health outcomes including IQ, among children born in the CPP study. Subjects were enrolled through 12 university-

affiliated medical clinics, with the centers contributing unequal numbers of subjects from 23 to 204. One of the hypotheses is that the total PCB level is related to the performance on the Weschler intelligence scale for children at 7 years of age. The total PCB was measured by analyzing the third-trimester blood serum specimens preserved from mothers in the CPP study. In our CPP data, total PCB was ascertained on all 850 subjects. We assume that they consist of the full study cohort. Some baseline characteristics of the CPP table are presented in Table 5.2.

To implement the probability dependent sampling, we set $a = 1$ so that the domain of total PCB is divided into 3 non-overlapping strata: $A_1 = (-\infty, 1.24]$, $A_2 = (1.24, 5.05]$, $A_3 = (5.05, \infty)$. A simple random sample of size 200 is selected from the full study cohort. We fit two separate logistic regression models with random intercept to predict $\hat{pr}(\text{PCB} \in A_1|\text{IQ}, Z)$ and $\hat{pr}(\text{PCB} \in A_3|\text{IQ}, Z)$, where Z is the vector of covariates available on all subjects in the full cohort. Specifically, Z includes standardized maternal education at birth of child, socioeconomic index, standardized maternal age at registration, race (black = 1), gender (female = 1), DDT, total cholesterol and triglycerides. Based on the predicted probabilities, we conduct supplementary sampling with $c = 85\%$ and select 100 subjects from both probabilistic tails. Using $i = 1, \ldots, 12$ to index medical clinics and j to index subject within the clinic, we postulate the following linear mixed effect model on IQ

$$\text{IQ}_{ij} = \beta_0 + \beta_1 \text{PCB}_{ij} + \beta_2 \text{Educ}_{ij} + \eta_i + \epsilon_{ij}, \tag{5.12}$$

where Educ is the standardized maternal education at birth of child. We also implemented the same model (5.12) on the data from (1) the full cohort and (2) a simple random sample of size 400. Analysis results are reported in Table 5.3. All three estimators identify a highly significant positive relationship between maternal education and IQ-score, that is, higher maternal education is significantly associated with higher IQ-score of the born child. On the other hand, none of them demonstrate a significant total PCB effect on IQ-score. The

corresponding P values were 0.8455, 0.6508 and 0.7848, respectively for full cohort, SRS and PDS estimators. Nonetheless, $\hat{\beta}_{PDS}$ has a smaller standard error compared to $\hat{\beta}_{SRS}$ and will result in a narrower confidence interval. For example, the 95% confidence interval for PDS estimator is $(2.00, 4.47)$, while the counterpart for SRS estimator is $(2.40, 5.18)$.

5.5. Discussion

Compared to the classical ODS, PDS does not require assuming a true underlying linear relationship between response and exposure, hence offers more flexibility. In this paper, we proposed a new random effect model for a cluster-based probability dependent sampling scheme. The implementation of PDS scheme involved drawing a first-stage simple random sample. Using this validation sample, we obtained the predicted probabilities of X falling in pre-specified lower and upper tails, then performed supplementary probability sampling. The estimation and inference procedures were based on a profile likelihood function. The procedure also properly accounted for the cluster-level heterogeneity. We observed improved efficiency over the estimator from a complete simple random sample in both simulation studies and real data analysis.

One potential drawback of our method lies in the added computational burden to address the cluster level random effect. In order to compute double integrals in (5.8), we applied the trapezoidal rule on a fairly fine grid of response Y and random effect η. The computing speed for this naive method was acceptable, because we do not need the numerical approximation to be extremely precise. We also considered other numerical integral methods like adaptive cubature. However, we failed to observe improved performance over the trapezoidal rule. Another approach to handle the numerical integral is Laplace approximation similar to the technique used in (Xu and Zhou 2012). By using Laplace approximation, each cluster-specific random effect η_i becomes an unknown parameter to be estimated. While this approximation may be more efficient with a small number of large

clusters (e.g. in CPP), the Newton-Raphson algorithm will be intractable when we have many small clusters.

While reviewing the literature, we noticed there was a lack of methodology for cluster-based ODS scheme. It is important to fill this gap. We can then further assess the performance of our cluster-based PDS estimator by comparing it with its ODS counterpart. PDS scheme also has the potential to be extended beyond the scope of continuous responses. One possible extension is to time-to-event data. PDS may be implemented in combination with the generalized case-cohort sampling design (Cai and Zeng 2007). Specifically, PDS can be embedded into the case-sampling stage of generalized case-cohort design. We can postulate a model for the failure times on the first stage simple random sample. We then obtain the predicted probabilities of a failure time is smaller or larger than a pre-specified threshold. We can sample the remaining cases based on the predicted probabilities and get a more representative second-stage supplementary sample, which may lead to improved statistical efficiency.

5.6. Proof of Theorems

Conditions

We required the following five regularity assumptions in the derivation of the asymptotic theories.

Assumption 5.6.1. *The log-density* $log\{f_\beta(Y|X,Z)\}$ *is twice continuously differentiable with respect to* β.

Assumption 5.6.2. *The proportion* n_k/n *is a fixed constant* $q_k \in (0,1)$ *for* $k = 0, 1, 3$.

Assumption 5.6.3. *The class of functions*

$$\mathcal{F} \equiv \left\{ \int f_\beta(Y|X,Z,\eta)d\eta, \frac{\partial^s}{\partial\beta^s} log\{\int f_\beta(Y|X,Z,\eta)d\eta\}, \frac{\partial^s}{\partial\beta^s} F_k(X,Z) : s = 0,1,2 \right\}$$

is a P-Donsker class and has an envelope function with finite second moment. The class of functions are indexed by ξ and parameters in ϕ.

Assumption 5.6.4. *The information matrix $-E[n^{-1}\partial^2 l(\xi)/\partial\xi\partial\xi^T]$ generated by likelihood function* (5.9) *is continuous in a neighborhood of true parameter ξ_0 and is non-singular at ξ_0.*

Assumption 5.6.5. *The estimated functions*

$$\frac{\partial^s}{\partial\beta^s}\hat{F}_k(X,Z) : s = 0,1,2$$

also belong to class $\mathcal{F}$ and $Pr[(Y,Z) : I\{\hat{\phi}_k(Y,Z) \geq c_k\} \to I\{\phi_k(Y,Z) \geq c_k\}] = 1$

Proof of Consistency

We compute the first-order derivatives of $n^{-1}\hat{l}(\xi)$ and evaluate them at the true value $\xi_0 = (\beta_0^T, \pi_1, \pi_3, 0, 0)^T$. There are three components

$$\begin{aligned}\frac{\partial n^{-1}\hat{l}(\xi)}{\partial\beta} &= n^{-1}\sum_{i=1}^{m}\frac{\partial}{\partial\beta} log\{\int\prod_{j=1}^{n_i} f_\beta(Y_{ij}|X_{ij},Z_{ij},\eta)d\eta\} \\ &\quad - n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\sum_{k=1,3} q_k\pi_k^{-1}\partial\hat{F}_k(X_{ij},Z_{ij})/\partial\beta}{\hat{\Delta}(X_{ij},Z_{ij})},\end{aligned} \tag{5.13}$$

and for $k = 1, 3$,

$$\frac{\partial n^{-1}\hat{l}(\xi)}{\partial \pi_k} = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{q_k\pi_k^{-2}\hat{F}_k(X_{ij}, Z_{ij})}{\hat{\Delta}(X_{ij}, Z_{ij})} - \frac{q_k}{\pi_k}, \tag{5.14}$$

$$\frac{\partial n^{-1}\hat{l}(\xi)}{\partial \nu_k} = -n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\hat{F}_k(X_{ij}, Z_{ij}) - \pi_k}{\hat{\Delta}(X_{ij}, Z_{ij})}. \tag{5.15}$$

By the Donsker property assumptions 5.6.3 and 5.6.5, we can apply the Glivenko-Canteli theorem respectively to (5.13), (5.14) and (5.15). We obtain

$$\left|\frac{\partial}{\partial \xi}n^{-1}\hat{l}(\xi) - \frac{\partial}{\partial \xi}E[n^{-1}\hat{l}(\xi)]\right| \xrightarrow{a.s.} 0.$$

Assumption 5.6.5 implies that $E[n^{-1}\hat{l}(\xi)] \to E[n^{-1}l(\xi)]$, then we have

$$\left|n^{-1}\frac{\partial}{\partial \xi}\hat{l}(\xi) - \frac{\partial}{\partial \xi}E[n^{-1}l(\xi)]\right| \xrightarrow{a.s.} 0.$$

Using similar arguments, we can show that

$$\left|n^{-1}\frac{\partial^2\hat{l}(\xi)}{\partial \xi \partial \xi^T} - \frac{\partial^2 E[n^{-1}l(\xi)]}{\partial \xi \partial \xi^T}\right| \xrightarrow{a.s.} 0,$$

for ξ in a neighborhood of the true value ξ_0.

We can use inverse mapping theorem arguments similar to those in Foutz (1977) to establish consistency of $\hat{\xi}$. The next step is to show that $n^{-1}\partial\hat{l}(\xi)/\partial\xi \to 0$. It suffices to show that $E[n^{-1}\partial l(\xi)/\partial\xi] \to 0$. To see that, replace the function $\hat{\phi}$ with ϕ in (5.13), (5.14)

and (5.15), then we have

$$\frac{\partial n^{-1}l(\xi)}{\partial\beta} = \left\{ n^{-1}\frac{\partial}{\partial\beta}\left[\sum_{i=1}^{m} log\{\int\prod_{j=1}^{n_i} f_\beta(Y_{ij}|X_{ij},Z_{ij},\eta)d\eta\} - \sum_{k=1,3} n_k log(\pi_k)\right]\right\}$$
$$- \left\{ n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\sum_{k=1,3} q_k\pi_k^{-1}\partial F_k(X_{ij},Z_{ij})/\partial\beta}{\Delta(X_{ij},Z_{ij})}\right\}, \tag{5.16}$$

and for $k = 1, 3$,

$$\frac{\partial n^{-1}l(\xi)}{\partial\pi_k} = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{q_k\pi_k^{-2}F_k(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})} - \frac{q_k}{\pi_k}, \tag{5.17}$$

$$\frac{\partial n^{-1}l(\xi)}{\partial\nu_k} = -n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{F_k(X_{ij},Z_{ij})-\pi_k}{\Delta(X_{ij},Z_{ij})}. \tag{5.18}$$

Using the fact that $\pi_k = E[F_k(X_{ij},Z_{ij})], k = 1, 3$, we substitute $F_k(X_{ij},Z_{ij})$ with π_k in (5.16), (5.17) and (5.18). It is obvious that the expectations of (5.17) and (5.18) converge to 0, respectively. The first bracket in (5.16) is essentially a 'valid' conditional log-likelihood function involving only β. 'Valid' here means we can estimate and make inferences on β solely based on the conditional log-likelihood function, even though it is not the most efficient approach. Therefore, the expectation of the quantity in the first bracket in (5.16) converges to 0. The expectation of the quantity in the second bracket also converges to 0, in view of the ODS fact

$$E\left[n^{-1}\sum_{i=1}^{n} g(Y_i,X_i,Z_i)\right] = q_0E[g(Y_i,X_i,Z_i)] + \sum_{k=1,3} q_kE[g(Y_i,X_i,Z_i)|\phi_k(Y,Z)\geq c_k].$$

Combining the results, we have shown that $n^{-1}\partial\hat{l}(\xi)/\partial\xi \to 0$ almost surely, that is, 0 belongs to the image of $n^{-1}\partial\hat{l}(\xi)/\partial\xi$ in any given neighborhood of the true value ξ_0 when n is

sufficiently large. By condition 5.6.4, $n^{-1}\partial^2\hat{l}(\xi)/\partial\xi\partial\xi^T$ is invertible in this neighborhood when n is sufficiently large. By the inverse mapping theorem, $n^{-1}\partial\hat{l}(\xi)/\partial\xi$ is invertible in any small neighborhood of the true value ξ_0. Therefore, we conclude that there is a solution $\hat{\xi}$ to $\partial\hat{l}(\xi)/\partial\xi = 0$ and

$$\hat{\xi} \xrightarrow{a.s.} \xi_0.$$

Proof of Asymptotic Normality

We add a term on both sides of $n^{-1}\partial\hat{l}(\hat{\xi})/\partial\xi = 0$ to obtain

$$n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi} - E\left[n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi}\right] = -E\left[n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi}\right].$$

The Taylor expansion of $E[n^{-1}\partial\hat{l}(\hat{\xi})/\partial\xi]$ around ξ_0 is

$$E\left[n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi}\right] = E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right] + E\left[n^{-1}\frac{\partial^2\hat{l}(\xi^*)}{\partial\xi\partial\xi^T}\right](\hat{\xi} - \xi_0),$$

where ξ^* is on the line segment between $\hat{\xi}$ and ξ_0. We substitute $E[n^{-1}\partial\hat{l}(\hat{\xi})/\partial\xi]$ on the right-hand side of the equation with its expansion and get

$$n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi} - E\left[n^{-1}\frac{\partial\hat{l}(\hat{\xi})}{\partial\xi}\right] = -E\left[n^{-1}\frac{\partial^2\hat{l}(\xi^*)}{\partial\xi\partial\xi^T}\right](\hat{\xi} - \xi_0) - E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right]. \qquad (5.19)$$

The left-hand side of (5.19) is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}U(Y_{ij}, X_{ij}, Z_{ij})$

where $U(Y_{ij}, X_{ij}, Z_{ij}) =$

$$\begin{pmatrix} \frac{\partial}{\partial\beta} log\{\int f_\beta(Y_{ij}|X_{ij}, Z_{ij}, \eta)h(\eta)d\eta\} - \frac{1}{\Delta(X_{ij},Z_{ij})} \sum_{k=1,3} q_k \pi_k^{-1} \frac{\partial F_k(X_{ij},Z_{ij})}{\partial\beta} \\ \frac{q_1\pi_1^{-2}F_1(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})} - \frac{q_1}{\pi_1} \\ \frac{q_3\pi_3^{-2}F_3(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})} - \frac{q_3}{\pi_3} \\ \frac{F_1(X_{ij},Z_{ij})-\pi_1}{\Delta(X_{ij},Z_{ij})} \\ \frac{F_3(X_{ij},Z_{ij})-\pi_3}{\Delta(X_{ij},Z_{ij})} \end{pmatrix}.$$

By the Donsker properties in assumptions 5.6.3 and 5.6.5, $n^{-1/2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} U(Y_{ij}, X_{ij}, Z_{ij})$ converges weakly to a Gaussian process.

For the second term of the right-hand side of (5.19), we have

$$E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right] = E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right] - 0 = E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right] - E\left[n^{-1}\frac{\partial l(\xi_0)}{\partial\xi}\right],$$

where the right-hand side quantity can be expressed as the mean of the summand

$$\begin{pmatrix} \sum_{k=1,3} q_k \pi_k^{-1} \left[\frac{\partial\hat{F}_k(X_{ij},Z_{ij})/\partial\beta}{\hat{\Delta}(X_{ij},Z_{ij})} - \frac{\partial F_k(X_{ij},Z_{ij})/\partial\beta}{\Delta(X_{ij},Z_{ij})}\right] \\ q_1\pi_1^{-2}\left(\frac{\hat{F}_1(X_{ij},Z_{ij})}{\hat{\Delta}(X_{ij},Z_{ij})} - \frac{F_1(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})}\right) \\ q_3\pi_3^{-2}\left(\frac{\hat{F}_3(X_{ij},Z_{ij})}{\hat{\Delta}(X_{ij},Z_{ij})} - \frac{F_3(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})}\right) \\ \frac{\hat{F}_1(X_{ij},Z_{ij})-\pi_1}{\hat{\Delta}(X_{ij},Z_{ij})} - \frac{F_1(X_{ij},Z_{ij})-\pi_1}{\Delta(X_{ij},Z_{ij})} \\ \frac{\hat{F}_3(X_{ij},Z_{ij})-\pi_3}{\hat{\Delta}(X_{ij},Z_{ij})} - \frac{F_3(X_{ij},Z_{ij})-\pi_1}{\Delta(X_{ij},Z_{ij})} \end{pmatrix}.$$

We can show that the second term of the right-hand side of (5.19) converges weakly to a zero mean Gaussian process. Specifically, by the Donsker assumption 5.6.3 and 5.6.5, both

$$n^{-1/2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{1}{\hat{\Delta}(X_{ij}, Z_{ij})} \sum_{k=1,3} q_k \pi_k^{-1} \frac{\partial\hat{F}_k(X_{ij}, Z_{ij})}{\partial\beta}$$

and

$$n^{-1/2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\Delta(X_{ij},Z_{ij})}\sum_{k=1,3}q_k\pi_k^{-1}\frac{\partial F_k(X_{ij},Z_{ij})}{\partial\beta}$$

converge in distribution to a Gaussian process. Therefore, the distribution of their difference is also Gaussian with mean 0. This is easily seen by replacing $\hat{\phi}$ with ϕ in the expression. Likewise, the other four components can be shown to converge to zero mean Gaussian processes, respectively. To summarize, we have

$$E\left[n^{-1}\frac{\partial\hat{l}(\xi_0)}{\partial\xi}\right]-E\left[n^{-1}\frac{\partial l(\xi_0)}{\partial\xi}\right]\sim\begin{pmatrix}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1,3}q_k\pi_k^{-1}Q_{1k}(Y_{ij},X_{ij},Z_{ij})\\ \sum_{i=1}^{m}\sum_{j=1}^{n_i}q_1\pi_1^{-2}Q_{21}(Y_{ij},X_{ij},Z_{ij})\\ \sum_{i=1}^{m}\sum_{j=1}^{n_i}q_3\pi_3^{-2}Q_{23}(Y_{ij},X_{ij},Z_{ij})\\ \sum_{i=1}^{m}\sum_{j=1}^{n_i}Q_{31}(Y_{ij},X_{ij},Z_{ij})\\ \sum_{i=1}^{m}\sum_{j=1}^{n_i}Q_{33}(Y_{ij},X_{ij},Z_{ij}),\end{pmatrix}.$$

The quantity on the right follows a zero mean Gaussian distribution and for $k=1,3$,

$$Q_{1k}(Y_{ij},X_{ij},Z_{ij})=\sum_{k=1,3}\left[\frac{\partial\hat{F}_k(X_{ij},Z_{ij})/\partial\beta}{\hat{\Delta}(X_{ij},Z_{ij})}-\frac{\partial F_k(X_{ij},Z_{ij})/\partial\beta}{\Delta(X_{ij},Z_{ij})}\right],$$

$$Q_{2k}(Y_{ij},X_{ij},Z_{ij})=\frac{\hat{F}_k(X_{ij},Z_{ij})}{\hat{\Delta}(X_{ij},Z_{ij})}-\frac{F_k(X_{ij},Z_{ij})}{\Delta(X_{ij},Z_{ij})},$$

and

$$Q_{3k}(Y_{ij},X_{ij},Z_{ij})=\frac{\hat{F}_k(X_{ij},Z_{ij})-\pi_k}{\hat{\Delta}(X_{ij},Z_{ij})}-\frac{F_k(X_{ij},Z_{ij})-\pi_k}{\Delta(X_{ij},Z_{ij})}.$$

According to assumption 5.6.5 and consistency of $\hat{\xi}$, the matrix in the first term on the right-hand side of (5.19) satisfies

$$E\left[n^{-1}\frac{\partial^2\hat{l}(\xi^*)}{\partial\xi\partial\xi^T}\right]\to E\left[n^{-1}\frac{\partial^2 l(\xi_0)}{\partial\xi\partial\xi^T}\right]=V(\xi_0).$$

Finally, we combine all results and obtain

$$
\begin{aligned}
&- V(\xi_0) \cdot n^{1/2}(\hat{\xi} - \xi_0) \\
=&n^{-1/2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[U(Y_{ij}, X_{ij}, Z_{ij}) + \begin{pmatrix} \sum_{k=1,3} q_k \pi_k^{-1} Q_{1k}(Y_{ij}, X_{ij}, Z_{ij}) \\ q_1 \pi_1^{-2} Q_{21}(Y_{ij}, X_{ij}, Z_{ij}) \\ q_3 \pi_3^{-2} Q_{23}(Y_{ij}, X_{ij}, Z_{ij}) \\ Q_{31}(Y_{ij}, X_{ij}, Z_{ij}) \\ Q_{33}(Y_{ij}, X_{ij}, Z_{ij}) \end{pmatrix} \right], \qquad (5.20)
\end{aligned}
$$

where (5.20) converge weakly to a Gaussian distribution with covariance matrix $\Omega(\xi_0)$. The asymptotic normality of $\hat{\xi}$ thus follows.

Table 5.1: *Simulation Results*

ICC	a	Estimator	β_1					β_2				
			Mean	ESD	ESE	CR	RE	Mean	ESD	ESE	CR	RE
0.33	N/A	$\hat{\beta}_F$	2.001	0.023	0.023	0.954	10.641	-0.500	0.015	0.016	0.960	11.797
		$\hat{\beta}_S$	2.004	0.076	0.076	0.948	1.000	-0.502	0.052	0.054	0.946	1.000
	1	$\hat{\beta}_{P_1}$	2.002	0.064	0.061	0.947	1.412	-0.499	0.041	0.041	0.947	1.601
		$\hat{\beta}_{P_2}$	2.001	0.057	0.061	0.960	1.753	-0.504	0.041	0.040	0.950	1.629
	1.5	$\hat{\beta}_{P_1}$	2.001	0.060	0.061	0.949	1.626	-0.503	0.041	0.041	0.941	1.648
		$\hat{\beta}_{P_2}$	2.004	0.065	0.061	0.938	1.360	-0.497	0.040	0.041	0.944	1.734
0.5	N/A	$\hat{\beta}_F$	2.000	0.023	0.023	0.956	12.126	-0.500	0.015	0.016	0.964	13.299
		$\hat{\beta}_S$	2.001	0.079	0.077	0.950	1.000	-0.500	0.055	0.055	0.954	1.000
	1	$\hat{\beta}_{P_1}$	2.002	0.074	0.071	0.947	1.130	-0.498	0.048	0.047	0.943	1.333
		$\hat{\beta}_{P_2}$	2.006	0.070	0.070	0.958	1.275	-0.502	0.048	0.047	0.934	1.356
	1.5	$\hat{\beta}_{P_1}$	2.004	0.067	0.071	0.955	1.390	-0.505	0.046	0.047	0.949	1.462
		$\hat{\beta}_{P_2}$	2.006	0.074	0.071	0.926	1.129	-0.497	0.046	0.047	0.950	1.470

NOTE: ESD, empirical standard deviation; ESE, average standard error estimator; CR, estimated standard error coverage rate of the nominal 95% confidence intervals. True values: $\beta_1 = 2$, $\beta_2 = -0.5$.

Table 5.2: *Baseline Characteristics of CPP Data*

Center Code	5	10	15	31	37	45
Sample Size	204	48	49	25	70	61
Total PCB: Lower Tertile	2.4	2.86	1.45	2.22	2.75	2.9
Total PCB: Upper Tertile	3.59	4.34	2.21	3.19	3.96	4.01
IQ: Mean (Std Dev)	102.98 (13.07)	107.9 (13.6)	85.31 (11.49)	90.96 (15.24)	99.23 (12.13)	88.66 (11.96)
Center Code	50	55	60	66	71	82
Sample Size	47	23	47	149	64	63
Total PCB: Lower Tertile	1.54	1.45	1.15	2.54	2.52	1.84
Total PCB: Upper Tertile	2.37	2.42	2	3.7	3.15	2.48
IQ: Mean (Std Dev)	104.79 (14.4)	86.13 (12.05)	95.55 (13.69)	90.06 (11.66)	96.42 (12.02)	90.05 (11.51)

Center codes: 5 - Boston Lying-in & Children's Hospital; 10 - Children's Hospital of Buffalo; 15 - Charity Hospital in New Orleans; 31 - Columbia University; 37 - Johns Hopkins University; 45 - Medical College of Virginia; 50 - University of Minnesota; 55 - New York Medical College; 60 - University of Oregon; 66 - University of Pennsylvania Hospital ; 71 - Providence Lying-in Hospital(Yale); 82 - University of Tennessee
Overall: sample size = 850, lower tertile of total PCB = 2.2, upper tertile of total PCB = 3.37, Mean(SD) of IQ = 95.65(14.26)

Table 5.3: *Analysis Results of CPP Data*

	$\hat{\beta}_{Full}$			$\hat{\beta}_{SRS}$			$\hat{\beta}_{PDS}$		
Effect	Estimate	Std Err	P Value	Estimate	Std Err	P Value	Estimate	Std Err	P Value
Total PCB	0.046	0.236	0.8455	-0.153	0.338	0.6508	0.077	0.282	0.7848
Education	3.638	0.452	$< .0001$	3.791	0.708	$< .0001$	3.237	0.631	$< .0001$

CHAPTER 6: SUMMARY AND FUTURE RESEARCH

In medical studies, correlated data may occur on various occasions. For example, one subject may experience multiple outcomes of interest that are correlated. On the other hand, subjects within the same cluster may be correlated because they share some similar characteristics. Proper methods are needed to analyze such data. This dissertation concentrates on multivariate statistical methods for correlated data in observation studies, possibly with biased sampling schemes.

The case-cohort design is widely used in large cohort studies when it is prohibitively costly to measure some exposures for all subjects in the full cohort, especially in studies where the disease rate is low. In Chapter 3, we have considered case-cohort designs with multiple disease outcomes. Our focus was on the marginal proportional hazard regression model in which the correlations among the failure times within each subject were considered as nuisance. In order to improve the statistical efficiency, we proposed a class of doubly-weighted estimators that can make better use of the covariate information in the subjects outside the case-cohort sample. The doubly-weighted estimator is also applicable to generalized case-cohort designs. We showed that our estimator was consistent and asymptotically normal. We observed improved statistical efficiency in simulation studies, with properly chosen second level weight functions. We also implemented our method to a data set from the Atherosclerosis Risk in Communities (ARIC) study.

In Chapter 4, we have considered the marginal structural Cox model for clusters of correlated failure time observations. This project was motivated by the growing popularity in observational electronic medical records (EMR) data. In EMR data, confounding by indication was usually a concern because treatments were unlikely to be assigned randomly.

Meanwhile, subjects in EMR data may form clusters within communities or clinics and the cluster level heterogeneity needed to be properly addressed. We formulated marginal structural Cox model and proved the consistency and asymptotic normality of the estimator. Simulation studies showed that marginal structural Cox model performed well by yielding unbiased estimate and satisfactory confidence interval coverage. The proposed method was implemented using a claim data assessing the effectiveness of the INSPIRIS home visiting health care program.

In Chapter 5, we studied the cluster-based probability-dependent sampling (PDS) design. PDS design is a biased sampling design for continuous responses. With PDS sampling, investigators can acquire a more informative biased sample (compared to a simple random sample of the same size), which may lead to more statically efficient estimators. Like the case in Chapter 4, subjects may form clusters in PDS designs. We proposed estimation and inference procedures that can address the cluster-level heterogeneity, based on a semiparametric profile likelihood function. The consistency and asymptotic normality of the cluster-based PDS estimator were proved using modern empirical process theory. In simulation studies, our method yielded more efficient estimators compared to linear mixed effect models on an SRS of the same size. We also applied the method to a data set from the Collaborative Perinatal Project.

The future work on the proposed methods in this dissertation includes:

First, in Chapter 3, our doubly-weighted estimator was based on Cox-type marginal proportional hazards model (Kang and Cai 2009). As an alternative to the proportional hazards model, we can extend doubly-weighted approach to marginal additive hazards model for (generalized) case-cohort studies with multiple disease outcomes (Kang et al. 2013). On the other hand, our method pertained to correlated events that do not compete. For example, in the ARIC study, it was possible for a subject to experience both CHD and incident stroke. However, there are cases when only one of the events of interest may

occur, e.g., deaths due to the disease of interest and deaths due to all other causes. We can consider statistical methods from a competing risks perspective (Sorensen and Andersen 2000) with doubly-weighting technique.

Second, the marginal structural Cox model we considered in Chapter 4 was based on the AG model (Andersen and Gill 1982), thus was ideal for terminal events such as death. As mortality information is rarely recorded in electronic claim databases, it will be important to extend our method to other common responses in such databases. One possible direction is to adapt our method based on the rate models by Pepe and Cai (1993) for recurrent events like hospitalization and emergency room visit. Another feasible extension is based on the proportional means regression model by Lin (2000a) for claim payments.

Third, to our best knowledge, there was a lack of methodology for cluster-based ODS design. It is critical to fill this gap. PDS scheme also has the potential to be extended beyond the scope of continuous responses. One possible extension is to time-to-event data. PDS may be implemented in combination with the generalized case-cohort sampling design (Cai and Zeng 2007). Specifically, PDS can be embedded into the case-sampling stage of generalized case-cohort design. We can postulate a model for the failure times on the first stage simple random sample. We then obtain the predicted probabilities of a failure time is smaller or larger than a pre-specified threshold. We can sample the remaining cases based on the predicted probabilities and get a more representative second-stage supplementary sample, which may lead to improved statistical efficiency.

BIBLIOGRAPHY

Andersen, P. K., R. D. Gill. 1982. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**(4) pp. 1100–1120.

Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* **59**(1) 19–35.

Austin, Peter C., Paul Grootendorst, Sharon-Lise T. Normand, Geoffrey M. Anderson. 2007. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. *Statistics in Medicine* **26**(4) 754–768.

Ballantyne, C. M., R. C. Hoogeveen, H. Bang, et al. 2005. Lipoprotein-associated phospholipase a2, high-sensitivity c-reactive protein, and risk for incident ischemic stroke in middle-aged men and women in the atherosclerosis risk in communities (aric) study. *Archives of Internal Medicine* **165**(21) 2479–2484.

Ballantyne, Christie M., Ron C. Hoogeveen, Heejung Bang, Josef Coresh, Aaron R. Folsom, Gerardo Heiss, A. R. Sharrett. 2004. Lipoprotein-associated phospholipase a2, high-sensitivity c-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the atherosclerosis risk in communities (aric) study. *Circulation* **109**(7) 837–842.

Barlow, William E. 1994. Robust variance estimation for the case-cohort design. *Biometrics* **50**(4) pp. 1064–1072.

Benson, Kjell, Arthur J. Hartz. 2000. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* **342**(25) 1878–1886.

Borgan, Ornulf, Bryan Langholz, SvenOve Samuelsen, Larry Goldstein, Janice Pogoda. 2000. Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6**(1) 39–58.

Breslow, Norman E., Jon A. Wellner. 2007. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics* **34**(1) 86–102.

Cai, Jianwen, Ross L. Prentice. 1995. Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**(1) 151–164.

Cai, Jianwen, Donglin Zeng. 2007. Power calculation for caseâĂŞcohort studies with nonrare events. *Biometrics* **63**(4) 1288–1295.

Chen, K., S-H Lo. 1999. Case-cohort and case-control analysis with cox's model. *Biometrika* **86**(4) 755–764.

Clayton, David, Jack Cuzick. 1985. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)* **148**(2) pp. 82–117.

Cole, Stephen R., Constantine E. Frangakis. 2009. The consistency statement in causal inference: A definition or an assumption? *Epidemiology* **20**(1).

Concato, John, Nirav Shah, Ralph I. Horwitz. 2000. Randomized, controlled trials, observational

studies, and the hierarchy of research designs. *New England Journal of Medicine* **342**(25) 1887–1892.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2) pp. 187–220.

Cox, D. R. 1975. Partial likelihood. *Biometrika* **62**(2) 269–276.

Danaei, Goodarz, Luis A. Garcia Rodriguez, Oscar Fernandez Cantero, Roger Logan, Miguel A. Hernan. 2013. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical methods in medical research* **22**(1) 70–96.

Ding, Jieli, Yanyan Liu, David B. Peden, Steven R. Kleeberger, Haibo Zhou. 2012. Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme. *Canadian Journal of Statistics* **40**(2) 282–303.

Ding, Jieli, Haibo Zhou, Yanyan Liu, Jianwen Cai, Matthew P. Longnecker. 2014. Estimating effect of environmental contaminants on women's subfecundity for the moba study data with an outcome-dependent sampling scheme. *Biostatistics* .

Fitzmaurice, Garrett, Nan Laird, James Ware. 2004. *Applied Longitudinal Analysis*. Wiley, Hoboken, New Jersey.

Foutz, Robert V. 1977. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72**(357) 147–148.

Gail, M. H., S. Wieand, S. Piantadosi. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**(3) 431–444.

Hájek, J. 1960. Limiting Distributions in Simple Random Sampling from a Finite Population. *Publications of Mathematical Institute of Hungarian Academy of Sciences, Series A* **5** 361–374.

Hernan, Miguel, Babette Brumback, James M. Robins. 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96**(454) 440–448.

Hernan, Miguel A., Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B. Michels, Walter C. Willett, JoAnn E. Manson, James M. Robins. 2008. Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* **19**(6) 766–779.

Hernan, Miguel A., James M. Robins. 2006. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health* **60**(7) 578–586.

Hernan, Miguel Angel, Babette Brumback, James M. Robins. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* **11**(5) pp. 561–570.

Hougaard, Philip. 1986. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**(2) 387–396.

Hougaard, Philip. 1995. Frailty models for survival data. *Lifetime Data Analysis* **1** 255–273.

Iezzoni, Lisa I., Arlene S. Ash, Michael Shwartz, Jennifer Daley, John S. Hughes, Yevgenia D. Mackiernan. 1995. Predicting who dies depends on how severity is measured: Implications for evaluating patient outcomes. *Annals of Internal Medicine* **123**(10) 763–770.

Ioannidis, JA, A Haidich, M Pappa, et al. 2001. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA: The Journal of the American Medical Association* **286**(7) 821–830.

Jha, Ashish K., Catherine M. DesRoches, Eric G. Campbell, Karen Donelan, Sowmya R. Rao, Timothy G. Ferris, Alexandra Shields, Sara Rosenbaum, David Blumenthal. 2009. Use of electronic health records in u.s. hospitals. *New England Journal of Medicine* **360**(16) 1628–1638.

Kalbfleisch, J. D., Ross L. Prentice. 2002. *The statistical analysis of failure time data*. John Wiley, Hoboken, N.J.

Kang, S., J. Cai. 2009. Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika* **96**(4) 887–901.

Kang, Sangwook, Jianwen Cai, Lloyd Chambless. 2013. Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the atherosclerosis risk in communities (aric) study. *Biostatistics* **14**(1) 28–41.

Kim, S., J. Cai, W. Lu. 2013. More efficient estimators for case-cohort studies. *Biometrika* .

Kulich, Michal, D. Y. Lin. 2004. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99**(467) 832–844.

Lee, E. W., L. J. Wei, D. A. Amato. 1992. Cox-type regression analyais for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art. J. P. Klein and P.K. Goel (eds.)* .

Lee, H. 2013. Marginal structural cox models with case-cohort sampling. *Dissertation* .

Lin, D. Y. 2000a. Proportional means regression for censored medical costs. *Biometrics* **56**(3) 775–778.

Lin, D. Y., Z. Ying. 1993. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* **88**(424) pp. 1341–1349.

Lin, DY. 2000b. On fitting cox's proportional hazards models to survey data. *Biometrika* **87**(1) 37–47.

Lu, Shou-En, Joanna H. Shih. 2006. Case-cohort designs and analysis for clustered failure time data. *Biometrics* **62**(4) 1138–1148.

Niswander, Kenneth R, Myron Gordon. 1972. *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. Saunders.

Owen, Art. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**(2) 237–249.

Owen, Art. 1990. Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18**(1) 90–120.

Pepe, Margaret Sullivan, Jianwen Cai. 1993. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**(423) 811–820.

Pine, Michael, Marija Norusis, Barbara Jones, Gary E. Rosenthal. 1997. Predictions of hospital mortality rates: A comparison of data sources. *Annals of Internal Medicine* **126**(5) 347–354.

Prentice, R. L. 1986. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**(1) 1–11.

Prentice, Ross, R. Pyke. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**(3) 403–411.

Qi, Lihong, C. Y. Wang, Ross L. Prentice. 2005. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **100**(472) 1250–1263.

Qin, Guoyou, Haibo Zhou. 2011. Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. *Biostatistics* **12**(3) 506–520.

Qin, Jing. 1993. Empirical likelihood in biased sample problems. *The Annals of Statistics* **21**(3) pp. 1182–1196.

Robins, James M. 1999. Association, causation, and marginal structural models. *Synthese* **121**(1/2) pp. 151–179.

Robins, James M., Miguel Angel Hernan, Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5) pp. 550–560.

Samuelsen, Sven OVE, Hallvard Anestad, Anders Skrondal. 2007. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics* **34**(1) 103–119.

Sasieni, Peter. 1993. Maximum weighted partial likelihood estimators for the cox model. *Journal of the American Statistical Association* **88**(421) pp. 144–152.

Schildcrout, Jonathan S., Shawn P. Garbett, Patrick J. Heagerty. 2013. Outcome vector dependent sampling with longitudinal continuous response data: Stratified sampling based on summary statistics. *Biometrics* **69**(2) 405–416.

Self, Steven G., Ross L. Prentice. 1988. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics* **16**(1) pp. 64–81.

Sorensen, P, PK Andersen. 2000. Competing risks analysis of the case-cohort design. *Biometrika* **87**(1) 49–59.

Spiekerman, C. F., D. Y. Lin. 1998. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* **93**(443) pp. 1164–1175.

Sturmer, Til, Michele Jonsson Funk, Charles Poole, M. A. Brookhart. 2011. Nonexperimental comparative effectiveness research using linked healthcare databases. *Epidemiology* **22**(3).

Tannen, Richard L, Mark G Weiner, Dawei Xie. 2009. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* .

Tsiatis, Anastasios A., Marie Davidian, Min Zhang, Xiaomin Lu. 2008. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine* **27**(23) 4658–4677.

van der Vaart, A. W. 1996. *Weak convergence and empirical processes*. Springer, New York.

Vardi, Y. 1982. Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**(2) pp. 616–620.

Vardi, Y. 1985. Empirical distributions in selection bias models. *The Annals of Statistics* **13**(1) pp. 178–203.

Wacholder, Sholom, Mitchell H. Gail, David Pee, Ron Brookmeyer. 1989. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika* **76**(1) pp. 117–123.

Wang, Xiaofei, Haibo Zhou. 2010. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics* **66**(2) 502–511.

Weaver, Mark A., Haibo Zhou. 2005. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100**(470) 459–469.

Wei, L. J., D. Y. Lin, L. Weissfeld. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**(408) pp. 1065–1073.

Xu, Wangli, Haibo Zhou. 2012. Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. *Biostatistics* **13**(4) 650–664.

Yamaguchi, Takuhiro, Yasuo Ohashi. 2004. Adjusting for differential proportions of second-line treatment in cancer clinical trials. part i: Structural nested models and marginal structural models to test and estimate treatment arm effects. *Statistics in medicine* **23**(13) 1991–2003.

Zhang, Min, Yanping Wang. 2012. Estimating treatment effects from a randomized clinical trial in the presence of a secondary treatment. *Biostatistics* **13**(4) 625–636.

Zhou, Haibo, Guoyou Qin, Matthew P. Longnecker. 2011a. A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal pcb exposure on cognitive function in children. *Biometrics* **67**(3) 876–885.

Zhou, Haibo, M. A. Weaver, J. Qin, M. P. Longnecker, M. C. Wang. 2002. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58**(2) 413–421.

Zhou, Haibo, Yuanshan Wu, Yanyan Liu, Jianwen Cai. 2011b. Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics* **12**(3) 521–534.

Zhou, Haibo, Wangli Xu, Donglin Zeng, Jianwen Cai. 2014. Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1) 197–215.

Zhou, Haibo, Jinhong You, Guoyou Qin, Matthew P. Longnecker. 2011c. A partially linear regression model for data from an outcome-dependent sampling design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **60**(4) 559–574.